Constructing Syntax-Based Distributional Semantic Models for Novel Languages

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik
der Universität Stuttgart zur Erlangung der Würde eines
Doktors der Philosophie (Dr. phil.) genehmigte Abhandlung

vorgelegt von

Jason Utt

aus Washington, USA

Hauptberichter:	Prof. Dr. Sebastian Padó
Mitberichter:	Prof. Dr. Alessandro Lenci
Tag der mündlichen Prüfung:	30.09.2019

Institut für Maschinelle Sprachverarbeitung
der Universität Stuttgart
2019

Ἐν ἀρχῇ ἦν ὁ Λόγος,
καὶ ὁ Λόγος ἦν πρὸς τὸν Θεόν,
καὶ Θεὸς ἦν ὁ Λόγος.

Abstract

Computational models of word meaning typically rely on large collections of text data in the language of interest. In the age of ever-increasing numbers of websites the text corpora needed for creating such distributional semantic models (DSMs) that are robust and high-coverage are becoming available in more and more languages. Among the more versatile DSMs are the structured DSMs (SDSMs) which are based on a more fine-grained view of what constitutes context: moving beyond simple neighboring words to include syntactic as well as other relational information. They thus allow relatedness estimates that are more relational–as opposed to topical–in nature as well as the modeling of the meaning of combinations of words via these syntactic relations. To construct SDSMs, however, simple text data alone does not suffice, instead requiring reliable and efficient parsing methods to obtain the syntactic structures, effectively limiting the number of languages for which such models can reliably be constructed.

This thesis explores methods of inducing structured distributional semantic models for novel languages and evaluates them on a battery of semantic tasks. I induce a monolingual SDSM from a moderately sized target language corpus as well as explore methods for doing away with any resource requirements beyond a simple bilingual lexicon to induce a cross-lingual SDSMs. Further, I show that it is possible to combine these two SDSM types to obtain high-coverage and high-accuracy multilingual models.

Zusammenfassung

Rechner-gestützte Modelle von Wortbedeutung bedürfen typischerweise umfangreiche Textdaten in der gewünschten Zielsprache. Heutzutage sorgt die ständig wachsende Anzahl von frei verfügbaren Webseiten dafür, dass die Erstellung solcher distributionellen semantischen Modellen (DSMs), welche robust und von hoher lexikalischen Abdeckung sind, in immer mehr Sprachen möglich wird. Zu den vielseitigsten DSMs gehören die strukturierten DSMs (SDSMs), welche den Kontextbegriff über einfache Nachbarworten auf syntaktische und andere Relationen ausdehnen. Dadurch erlauben sie Ähnlichkeitsvorhersagen, die über die thematischen Bedeutungsaspekte eines Wortes, oder gar einer syntaktischen Verknüpfung von Wörtern, hinaus auch die relationaler Natur einbeziehen. Textdaten alleine reichen jedoch nicht aus, um SDSMs zu konstruieren. Es werden zuverlässige und effiziente Parser in der Zielsprache benötigt, um die syntaktischen Analysen zu erhalten; was zur Folge hat, dass momentan leider nur wenige Sprachen von solchen Modellen profitieren können.

Diese Dissertation untersucht Verfahren, die es erlauben, für neue Sprachen strukturierte distributionelle semantische Modelle zu erzeugen und testet diese auf einer Reihe von semantischen Aufgaben. Es wird zunächst ein monolinguales SDSM von einem zielsprachigen Textcorpus mittler Größe erzeugt; werden Methoden ermittelt, mit denen man ausschließlich mithilfe eines einfachen bilingualen Lexikons ein cross-linguales SDSM. Weiter wird aufgezeigt, wie diese zwei SDSM-Typen verknüpft werden können, um ein multilinguales Modell zu erhalten, welches die Vorteile beider Eingabemodelle behält und somit hohe Abdeckungsraten mit genauen Vorhersagen aufweist.

Acknowledgments

First of all I would like to thank my advisor Sebastian Padó for all the support and encouragement along the path toward completing this project. Whether it came in the form of wise counsel, a bad joke or delicious baked goods his input was always welcome.

All of my work would obviously not have been possible without financial support which I enjoyed thanks to the German Research Foundation's funding of projects D6 and D10 in SFB 732.

For the discussions, inspiration and cooperation on many topics I would like to express my gratitude to Sylvia Springorum, Christian Scheible, Diego Frassinelli, Alessandra Zarcone, Britta Zeller, Jan Šnajder, Laura Aina, Sarah Hemmen, Paweł Müller, Alok Kothari, Abhijeet Gupta and Max Kisselew.

Last but certainly not least a hearty thank you to all my colleagues at the IMS – working alongside you made even our drab 'new' building feel as cozy as our old home in Azenbergstraße. Well, almost.

Contents

III Evaluating and Combining SDSMs

IV Conclusion

List of Figures

List of Tables

Part I.

Introduction and Background

Chapter 1.

Introduction

An area of active research that is central to the field of computational lexical semantics is the question of what type of representation captures the semantic properties of words best while still remaining computationally tractable. The class of vector space models (VSMs) arguably represent the most successful framework for semantic representation in the last couple of decades (Turney and Pantel, 2010). At the heart of a VSM model is the vector space: a matrix in which each word in the vocabulary is represented by a – potentially high-dimensional – vector.

Vector space models. The appeal of dealing with vector representations is that they inherently contain a notion of **similarity** qua spatial proximity. The first VSMs were term-document matrices used for information retrieval (Salton, 1989): In this setting, the rows would correspond to terms to be used as indices into the documents which were listed as columns and the cells contain the frequency of the given term in that document. This provides a dual view of the resulting space: Both documents and terms can be compared with respect to their shared information. When comparing documents those that contain a number of terms in similar proportions points to an implicit semantic similarity between them. Consider, e.g. two e-commerce platforms that might sell different products but their websites both contain the terms *purchase, order, cart, checkout* or *offer*. This situation allows us to compare the terms themselves in an analogous fashion: When it is observed that they co-occur across a large number of documents it seems reasonable to assume they are related.

Distributional models. Going beyond term-document matrices, a common form of VSM is the **distributional semantic model** (DSM) observes words in context in a text corpus and encodes the distributional information of these co-occurrences between word and context in the vectors.

The reason for considering such models semantic derives from the **distributional hypothesis**, based on considerations of Wittgenstein (1953); Harris (1954) and Firth (1957) which put in terms inspired by the famous quotation of Aesop states:

> *You shall know a word by the company it keeps.*

Distributional models operationalize this notion explicitly by observing words in a large number of contexts and learning the statistical associations between words and these contexts. The association functions can vary from simple co-occurrence counts – i.e. words seen in a context x n number of times will be assigned an association value n for that context x – or applying monotonic functions to those counts, e.g. logarithmic, $\log(n)$, or square root, $\sqrt{n}$, to more complex transformations using the whole vectors space, e.g. point-wise mutual information (Lowe, 2001). The resulting association scores of the target word with its observed context in a corpus are the components of the **word vector**. In their simplest form, the dimensions of the resulting vector space are words taken as atomic context elements.

	shoot	*eat*	*grass*
hunter	20	10	5
deer	5	10	10
bullet	10	0	1

Table 1.1.: Example space representing the words *hunter, deer* and *bullet*.

Given a vector space such as in Table 1.1, we can calculate similarities based on the shared contexts (cf. Chapter 2 for more details), using a standard cosine similarity measure, we obtain:

$$sim(deer, hunter) = 0.73$$
$$sim(deer, bullet) = 0.4$$
$$sim(bullet, hunter) = 0.89$$

i.e. a higher similarity between *deer* and *hunter* than should be desired.

Structured distributional semantic models. The issue at hand is obvious: Using only the words in the targets' contexts into account, there is no way of distinguishing typical subjects of a verb *shoot* such as *hunter* from its objects, e.g. *deer*, as the encoding of the context dimensions only tells us that they cooccur frequently as words, i.e. without relational information.

Structured distributional semantic models (SDSMs) (Lin, 1998; Padó and Lapata, 2007; Baroni and Lenci, 2010) are designed to alleviate this shortcoming by increasing the specificity of the context dimensions:

	$\langle obj\ shoot \rangle$	$\langle subj\ shoot \rangle$	$\langle eat\ grass \rangle$	$\langle subj\ eat \rangle$	$\langle obj\ eat \rangle$
hunter	1	19	0	10	0
deer	5	0	10	5	5
bullet	8	2	0	0	0

Now we have access to relational information, for instance that hunters have been more commonly seen as subjects of the verb *shoot* than objects – luckily for them – and vice versa for deer.

Here, the concept of structured contexts is first one of grammatical structure; we are encoding the relations in which the target word was seen with a particular verb, e.g. *hunter* was seen as the object of *shoot* than it was as its subject. But this notion can be more general than, we also include structures such as a subject-verb-object triple $\langle deer\ \text{EAT}\ grass \rangle$ without explicit labeling of grammatical functions.

Now we are able to identify finer grained types of contexts which allow for the commonsense distinctions between *hunter* and *deer*:

$$sim(deer, hunter) = 0.19$$
$$sim(deer, bullet) = 0.37$$
$$sim(bullet, hunter) = 0.26$$

However, we still must deal with the matter of direct objects of *shoot* being either the target or the projectile, hence the high similarity between *deer* and *bullet*.[1] The increase in informativeness also incurs a cost, namely, an increase in the sparsity in the vectors.

The variety of possibilities for defining the actual content or structure of these more informative features[2] gives rise to distributional models of varying computational and conceptual complexity. In a commensurate fashion, there are numerous fields of applicability of such models, including word sense disambiguation (McCarthy, Koeling, Weeds, and Carroll, 2004), the representation of selectional preferences (Erk, Padó, and Padó, 2010), verb class induction (Merlo and Stevenson, 2001; Schulte im Walde, 2006), analogical reasoning (Turney, 2006), or alternation discovery (Joanis, Stevenson, and James, 2006). As this array of applications contains quite distinct tasks, there are correspondingly many different types of semantic vector spaces. In cases of such traditional tasks as word sense disambiguation, we quantify the similarity or relatedness[3] between words themselves. By contrast, in the case of analogical reasoning, we need to compare pairs of words; whereas for alternation discovery we need to be able to compare verbal argument positions.

Distributional Memory (DM) – which is described in detail in Chapter 3 – encodes both syntactic and lexical context. Its English models have been shown (Baroni and Lenci, 2010) to be flexible and robust general-purpose SDSMs. There are, however, significant resource requirements for building such a model – a very large and accurately parsed corpus – which is not available for most languages.

Contributions. In this thesis, I describe and evaluate methods to build accurate and reliable SDSMs for novel languages:

[1] At the same time, we are necessarily confronted with lexical ambiguity, e.g. of words such as *graze*, of which both *deer* and *bullet* are highly suitable subjects. While lexical ambiguity is a challenge in all models of lexical semantics and will not be addressed in particular in this work.

[2] Popular choices include: dependency relations or paths (Lin, 1998; Padó and Lapata, 2007), subcategorization frames (Schulte im Walde, 2006) or semantic role labels (Sayeed, Demberg, and Shkadzko, 2015).

[3] Words are considered semantically similar when they are more or less synonymous, e.g. *car/vehicle*, whereas the class of semantically related words covers more distant pairs such as *dog/cat* or *car/traffic*.

1. monolingually, when parsed corpus data is available;
2. cross-lingually, by leveraging the existing English DM; and
3. by combining the SDSMs obtained by the above two approaches.

The evaluations are carried out on a variety of tasks ranging from simple lexical comparison to multi-argument composition tasks.

Structure of the thesis. This thesis is structured into four parts. The current part (Part I) introduces the themes of this thesis, sets distributional models into their historical and conceptual context (Chapter 2) and presents the models, datasets and tasks that are used (Chapter 3).

Part II covers the induction methods for mono- and cross-lingual DMs. We describe the methods used to induce German DM supplementing a qualitative analysis in Chapter 4 and in Chapter 5 we formulate the translation algorithm of a source-language DM into a target-language DM, with special attention given to to handling lexical ambiguity and translational variance.

In Part III, in depth evaluations are conducted on our monolingual and cross-lingual DMs (Chapter 6) and then combination schemes are tested, to leverage overcome the shortcomings of either model type (Chapter 7).

Finally, we draw conclusions and make suggestions as to future work in the area of constructing and testing DMs as an instance of SDSMs.

Previous Publications. A portion of the work presented here has previously been published.

- Monolingual induction and evaluation of a German DM:

  Padó, S. and J. Utt (2012). A distributional memory for German. In *Proceedings of the KONVENS 2012 workshop on recent developments and applications of lexical-semantic resources*, pp. 462–470.

  Zarcone, A., J. Utt, and S. Padó (2012). Modeling covert event retrieval in logical metonymy: Probabilistic and distributional accounts. In *Proceedings*

of the 3rd Workshop on Cognitive Modeling and Computational Linguistics, CMCL '12, pp. 70–79.

- Induction and evaluation of cross-lingual and combined DMs:

  Utt, J. and S. Padó (2014). Crosslingual and multilingual construction of syntax-based vector space models. *Transactions of the Association for Computational Linguistics 2*, 245–258.

- Experiments on Croatian:

  Padó, S., J. Šnajder, J. Utt, and B. Zeller (2016). Smoothing syntax-based semantic spaces: Let the winner take it all. In *Proceedings of KONVENS*, pp. 186–191.

The experiments presented therein were performed by the present author.

In this work, I embed these previously published results in a wider range of experiments on which the DMs are tested as well as more in-depth analyses of the models' performance. Conforming to the standard style in academic writing, I will employ the personal form of 'we' throughout this thesis.

Notational Conventions.

DSM	distributional semantic model	SL	source language
SDSM	structured DSM	TL	target language
DM	Distributional Memory	ML	monolingual
cos sim	cosine similarity	XL	cross-lingual
NLP	natural language processing	POS	part of speech
$\mathcal{M}$	model	$\mathfrak{M}$	multilingual model
$\pi_{\mathcal{M}}$	model prediction function	$\sigma_{\mathcal{M}}$	model scoring function
$\langle w_1\ w_2 \rangle$	word pair	$\langle w_1\ \text{L}\ w_2 \rangle$	word-link-word triple

E.g. the equation:

$$\pi_{\mathcal{M}}(\langle w_i\ w_j \rangle) \stackrel{\text{def}}{=} max_{\langle w_i\ \text{L}\ w_j \rangle \in \mathcal{M}} \left(\sigma(\langle w_i\ \text{L}\ w_j \rangle)\right)$$

defines the predictions of model $\mathcal{M}$ for a word pair $\langle w_i\ w_j \rangle$ as the maximum score over all w_i-L-w_j triples in $\mathcal{M}$.

Chapter 2.

Background

2.1. Formal representations of lexical meaning

Lexical semantic models have the goal of representing the meaning of words in a manner amenable to automated processing. Past approaches to formalizing the semantic content of linguistic forms had domains ranging from human languages to mathematical theories; an example of the latter being the program of axiomatization and verification – commonly referred to as metamathematics or the philosophy of mathematics – at the turn of the 20th century.

The attempts at systematizing what we mean with our speech – or our mathematical formulations – were based on truth theory: if two forms had the same conditions for being true, they had the same meaning. Montague (1970)'s revolutionary development of translation procedures from syntactic analyses of English into formulae of first-order logic, as well as later frameworks such as Discourse Representation Theory (Kamp and Reyle, 1993) which followed the same path, all had as the basis for their representations the notion that inferences were of prime importance.

While the utility of having a strong emphasis on logical inference is evident within these frameworks which focus on the composition of intra- to inter-sentential meaning to a larger, logical whole, such systems did not deal well with issues such as lexical ambiguity or vagueness of concepts (Bos and Markert, 2005). In addition, any definitional relation in meaning between words had to be hard-coded using **meaning postulates** (e.g. predicates such as: $\mathit{bachelor}(x) \stackrel{\text{def}}{=} \neg \mathit{married}(x)\ \&\ \mathit{male}(x)$) and **semantic primitives** (i.e. the fundamental predicates not derived from others,

e.g. $male(\cdot)$), otherwise the inference engine might infer inconsistent meanings – such as that someone who is a bachelor just went on his honeymoon.

The total set of facts and rules that can be applied to facts and other rules in order to generate new facts make up a **knowledge base**. Then the choice of what the semantic primitives expressed in the knowledge base are – and the work of actually encoding all semantic relationships that are to be axiomatic and not emergent – becomes a nearly intractable task, requiring care to not give rise to an inconsistent logical system.

Computational resources. One project to create and maintain a large-coverage knowledge base that provides access to the types of useful and interesting inferences is Cyc (Lenat, Guha, Pittman, Pratt, and Shepherd, 1990) which aims at being a global-scale expert system, based on a classical knowledge base approach.[1]

Lexical databases such as WORDNET (Miller, Beckwith, Fellbaum, Gross, and Miller, 1990; Beckwith, Fellbaum, Gross, and Miller, 1991; Miller and Fellbaum, 1992) and FRAMENET (Baker, Fillmore, and Lowe, 1998a,b,c) are also widely used in automated systems for processing meaning. WORDNET maps words to their distinct meanings which are organized hierarchically via a partial order.[2] With its emphasis on covering lexical relations, its design is based on psycholinguistic theories of how lexical items are stored and retrieved by the brain (Miller, 1986).

FRAMENET on the other hand starts from a semantic theory (Fillmore, 1976) according to which the understanding of lexical items requires knowledge of related other lexical items (so-called frame elements, e.g. the word *buyer* engendering a prerequisite conceptualization of the good(s) being bought, the seller, price etc.) into a frame. This means there is an attempt to relate words to some world knowledge, i.e., a formalization of how we expect things in the world to be or how we typically act in it. The frames in FRAMENET are organized into clusters around verbs which share frame elements, i.e. typical fillers for semantic roles evoked by

[1] The approach taken for Cyc was questioned by Marvin Minsky, the founder of artificial intelligence: while having seen decades' worth of work invested still lacks simple word knowledge that barely verbal humans have (Baard, 2003).

[2] Such a format can help for instance in textual entailment tasks (Lan and Jiang, 2018) which is similar in nature to inference tasks.

that frame. This represents a reliance on a set of semantic primitives.

2.2. Distributional models of meaning

Classification of DSMs. In more recent years, vector space models (Turney and Pantel, 2010) have become the de-facto standard among models in computational lexical semantics in NLP. The strengths of distributional models in contrast to those in the tradition of formal semantics lie in their ability to capture **gradation** in meaning (Erk and McCarthy, 2009; Basili and Pennacchiotti, 2010) and uncover latent relationships in usage (Landauer and Dumais, 1997a), both in production and perception, which has been considered to be the foundation for meaning (Lakoff and Johnson, 1980; Barsalou, 2008). Multi-modal distributional models (Bruni, Tran, and Baroni, 2014) which capture not only textual but also visual and even auditory (Kiela and Clark, 2015) features are the best examples of perceptual grounding of the features making up the representational space. For example, words can obviously have multiple unrelated meanings in the case of *[river] bank – [financial] bank*, but more often there are slight shifts in meaning that are typically only captured in discrete lists while in reality the meanings are clearly related:

John's paper won an award – I made 5 copies of the paper for the reading group

At the same time, such models can be obtained essentially 'for free' in an unsupervised fashion from large text corpora and without investing in any particular lexico-semantic or cognitive linguistic theory as lexical resources such as those summarized above do.

It is instructive to investigate how distributional models are to be viewed in terms of the classification suggested by Winograd (1978) which introduces three main aspects for characterizing approaches to formally represent linguistic meaning: an **ontological**, a **logical** and a **relational** aspect.

Ontological models use **symbol structures**, i.e. they are composed of discrete, symbols whose structured relationships are hard-coded or learned. The symbols – or the relationships themselves – are, depending on the theory, seen as having a strong or weak equivalence to psychological entities and processes. Logical

models, as outlined above, are primarily concerned with allowing the operation of logical inference or **implication** to be performed. Finally, the relational aspects of semantic models are present when terms or concepts are described directly via their relationship with others. Such relationships could be definitional or descriptive in nature and can, as is the case with most manually compiled dictionaries, even be circular.

As an overall operational approach[3], DSMs can be viewed as **relational** and weakly **ontological models**. They are relational in that target words are defined using their linguistic context in a definitional or descriptive manner. A representation system viewed from the ontological perspective will be concerned with implementing a system of symbols whose design depends on one's view on whence the symbols and their organization arise. We consider this as a valid view of what occurs in distributional models since the word-context association scores which make up such models directly reflect the words' usage in the world.[4] These symbols can themselves be **linguistic**, i.e. the system's meta-language is a sub- or superset of the object language[5]; **psychological**, the symbols and structures correspond to – or at least account for important aspects of – psychological entities or processes; or purely **theoretical**, such as neutrinos in particle physics being postulated before being observed to help 'fill in the gaps' of the prevalent theory. This results in the overview given in Table 2.1.

Properties of DSMs. DSMs can thus be viewed as linguistically based[7] and psychologically motivated: It is possible to inspect the models to make sense of semantic neighborhoods. We can also inspect those contexts that are more representative for a group of words' vectors, and make determinations as to whether those dimensions as more or less suited to some given or intuited psychological concepts.

[3] I.e. the models are primarily concerned with accounting for or predicting aspects of directly observable language use.

[4] Further, there is a latent organization or association between words as well as contexts, which is leveraged to reduce noise using dimensionality reduction methods (cf. below).

[5] That is, the language being studied.

[6] In terms of lexical entailment or combination with formal frameworks as described above.

[7] Words form the basic building blocks of the spaces' context dimensions.

Name	Description	Ontol.	Log.	Rel.
Dictionary	Definitions, descriptions			✓
Ontology	Taxonomy/graph (relations as connections, e.g. `is-a`)	✓	✓	✓
Feature norms	Descriptions, associative			✓
FrameNet	Linked definitions using semantic primitives	✓	(✓)	✓
DSMs	Weighted associations between words and contexts	(✓)	(✓[6])	✓

Table 2.1.: Organizational and representational forms for lexical information.

More importantly, we are able to predict psychological realities such as:

- correlates of **processing difficulty** (Chater and Manning, 2006; Mitchell, Lapata, Demberg, and Keller, 2010),
- **priming effects** (Padó and Lapata, 2007), or
- accounting for the **interpretation of metonymic constructions** (Zarcone, Utt, and Padó, 2012; Zarcone, Lenci, Padó, and Utt, 2013)[8]

Regarding their logical nature, one would have to show their ability to support entailment or inferences. Here we can see that lexical entailment – or hypernymy/hyponymy relation detection – has been successfully performed in purely distributional models (Roller, Erk, and Boleda, 2014; Roller and Erk, 2016; Shwartz, Goldberg, and Dagan, 2016; Chang, Wang, Vilnis, and McCallum, 2018).

Furthermore, research has investigated the promise of combining formal models with distributional ones to have both a strong entailment mechanism via symbolic structures, while at the same time maintaining a 'graded' notion of semantic relatedness from a vector space. By including distributional information, Lewis and Steedman (2013) were able to significantly boost recall while maintaining high accuracy in a wide-coverage question answering task. Beltagy, Roller, Cheng, Erk, and Mooney (2016) leveraged a distributional model to build a textual entailment

[8]Cf. also Section 6.4.

database and found that the lexical distributional representations helped capture complex paraphrases, such as *teenage* → *in teens*, *ride bike* → *biker*, *young lady* → *teenage girl*, which helped to boost performance, while using handcrafted lexical relations such synonymy, hypernymy and antonymy read off the WORDNET hierarchy led to a decrease in performance.

Additionally, while not being committed to any particular theory of meaning, it has been argued in the area of cognitive science that VSMs represent psycholinguistically plausible models (Landauer, McNamara, Dennis, and Kintsch, 2007; Turney and Pantel, 2010).

While resource types such as those listed in Table 2.1 have been explored along with attempts to automate their construction (Maedche and Staab, 2000; Westerhout, 2009; Navigli and Velardi, 2010; Navigli and Ponzetto, 2012), typical approaches for obtaining these resources rely on **manual compilation** or at least use some supervised methods. The main drawback of which, as opposed to most distributional methods, is their time- and labor-intensive nature.

Semantic relatedness. One straightforward use of a vector-based model is to estimate semantic relatedness between words. As Resnik (1995); Budanitsky and Hirst (2006) point out, there is an important distinction to be drawn between semantic similarity and semantic relatedness. In terms of a taxonomy, similarity can be captured in terms of local `is-a` relationships between words (Rada, Mili, Bicknell, and Blettner, 1989) (à la *truck – tricycle*), whereas the more general term relatedness covers more associative types of links in meaning and potentially word knowledge (à la *truck – mechanic*) provided by the vectors' relationships in space. Since vectors are geometric objects, they have **direction** and – if the vectors are not length-normalized – a distinct **locality** via their different lengths and can thus be compared in terms of the angle between them or their proximity. A corollary means that semantic relatedness so defined is **graded** in nature. It is instructive to consider the extreme cases of the possible spatial relationship of vectors. If the vectors are approximately identical or parallel, they share the same contexts in the same proportions and this suggests the words would have a high degree of substitutability.[9]

[9]Consider e.g. a similar scenario in the setting of language models: An n-gram language model assigning high probabilities to two words across a number of different preceding

The case of no shared contexts and – thus of semantic complementarity – would mean **orthogonality** in the vector space. Finally, it must be noted that there are multiple ways in which the proximity or distance in the space can be interpreted or operationalized within a theory – to a large degree depending on the definition of context – leading us to simplify the terminology and simply speak of **similarity**.

Formal description of VSM. In its simplest form, a vector space model of word meaning, each word w_i $(1 \leq i \leq T)$ in the set of T vocabulary terms is represented as a vector in D-dimensional space:

$$\vec{w}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iD} \end{pmatrix}$$

The values which make up the vector x_{ij} – its components – are the association scores[10] between the words w_i and their possible contexts c_j:

$$\vec{w}_i = \sum_i \mathit{assoc}(w_i, c_j) \times \vec{c}_j,$$

where the vectors $\vec{c}_j$ ($j \in \{1, ..., D\}$) definitionally form an orthogonal basis of $\mathbb{R}^D$. Thus any VSM can be stored as a matrix $X \in \mathbb{R}^{T \times D}$. A simple initial association score is the raw frequency of seeing a word in a context: $\mathit{assoc}(w_i, c_j) = \text{freq}_{ij}$.

Once all frequencies have been collected, reweighting schemes are commonly employed to discount overly common words or pairs. In the case of document-term spaces[11], tf-idf is the standard reweighting scheme which penalizes terms that occur in a large number of documents:

$$\begin{aligned} \mathit{assoc}(w_i, d_j) &= \text{term frequency} \times \text{inverse document frequency} \\ &= \text{freq}_{ij} \times \log \left(\frac{\#\,\text{documents}}{\#\,\text{documents containing}\, w_i} \right) \end{aligned}$$

was shown to improve results in information retrieval tasks (Salton and Buckley,

contexts would suggest both an amount of syntactic as well as high semantic overlap with respect to the model.

[10] The domain for the association values must be a field, and we use the most commonly chosen field, the real numbers $\mathbb{R}$ (Fischer, 2003).

[11] In such spaces, the document in which a term occurs in considered its context.

1988).

For word spaces, *PMI* (Fano, 1961) is a probabilistic reweighting method which quantifies the degree to which the joint probability of the word and context occurring exceeds the probability of them co-occurring by chance. With joint probability $p_{ij} = \frac{\text{freq}_{ij}}{\sum_{i,j} \text{freq}_{ij}}$ and marginal probabilities $p_{i*} = \frac{\sum_j \text{freq}_{ij}}{\sum_{i,j} \text{freq}_{ij}}$ and $p_{*j} = \frac{\sum_i \text{freq}_{ij}}{\sum_{i,j} \text{freq}_{ij}}$:

$$PMI = \log\left(\frac{p_{ij}}{p_{i*} \cdot p_{*j}}\right),$$

which is greater than 0 when the cooccurrences are more frequent than chance. It was used in the setting of association strength between words (Church and Hanks, 1990) to overcome the subjectivity of the elicitation studies in psycholinguistics where researchers were interested in word association norms (Palermo and Jenkins, 1964).

A dimensionality reduction algorithm $\mathfrak{R}$ will map X onto X' in which the terms are maintained but the number of dimensions has been significantly reduced such as Latent Semantic Analysis (LSA: Landauer and Dumais, 1997b):

$$\mathfrak{R} : \mathbb{R}^{T \times D} \rightarrow \mathbb{R}^{T \times D'}, \text{ where } D' \ll D.$$

One importance consequence of dimensionality reduction is that we can lose direct correspondence between dimensions and contexts. In most cases, however, $\mathfrak{R}$ is a matrix meaning the dimensions in D' can be identified as linear combinations of context words in D.

Similarity measures. While a number of methods for comparing vectors have been experimented with (Weeds, Weir, and McCarthy, 2004), the standard metrics euclidean distance and cosine similarity are both geometric in nature.

$$\textit{cos sim}(w_1, w_2) = \cos(\measuredangle(\vec{w}_1, \vec{w}_2)) = \frac{\sum_j x_{1j} \cdot x_{2j}}{|\vec{w}_1| \cdot |\vec{w}_2|}$$

$$\text{eucl}(w_1, w_2) = |\vec{w}_1 - \vec{w}_2| = \sqrt{\sum_j (x_{1j} - x_{2j})^2}$$

Figure 2.1 illustrates the relationship between the euclidean distance between two vectors and the angle between the vectors. The cosine similarity of two vectors is 0 when the vectors are orthogonal to one another. Orthogonality means that in all dimensions where either vector has a non-zero component, the other vector must have a zero-valued component. This does not preclude the vectors from sharing many zero-valued components; similar to the case of distance: Distance increases when there is a difference in any of the vectors' corresponding components.

Obviously, distance measures will have an inverse relationship to similarity measures. For length-normalized vectors, euclidean distance and cosine similarity give exactly inverse rankings, i.e. ranking a list of word pairs by their distances yields the exact opposite ordering than if they were ordered by cosine similarity.

While most commonly used similarity measures are symmetric, a number of asymmetric measures have been used to model directed relationships between words, such as hyponymy/hypernymy to model lexical entailment relations (Szpektor and Dagan, 2008; Kotlerman, Dagan, Szpektor, and Zhitomirsky-Geffet, 2009) or identify collocations (Michelbacher, Evert, and Schütze, 2007).

2.3. Bag of Words (BOW) Models

The simplest family of distributional models use a **bag-of-words assumption**, i.e. that the necessary context information can treat the target words' context as a figurative bag of words. This means the models make no distinction with regards to the relationships between or order of the words in the sentence. A context word is either in a target word's context or not.

After some standard NLP pre-processing steps (tokenization and lemmatization), we will have bare word forms as targets and contexts. One common way of reducing noise in the space is to filter out very common and semantically poor words such as function words (e.g. *the*, *of*) or pronouns (e.g. *they*, *it*) contained in a list of **stop words**.

Zipf (1949) discovered that a word's frequency follows an inverse power law with regards to its frequency rank. This shows an important issue in distributional modeling: **data sparsity**. We require many examples of a word in context to

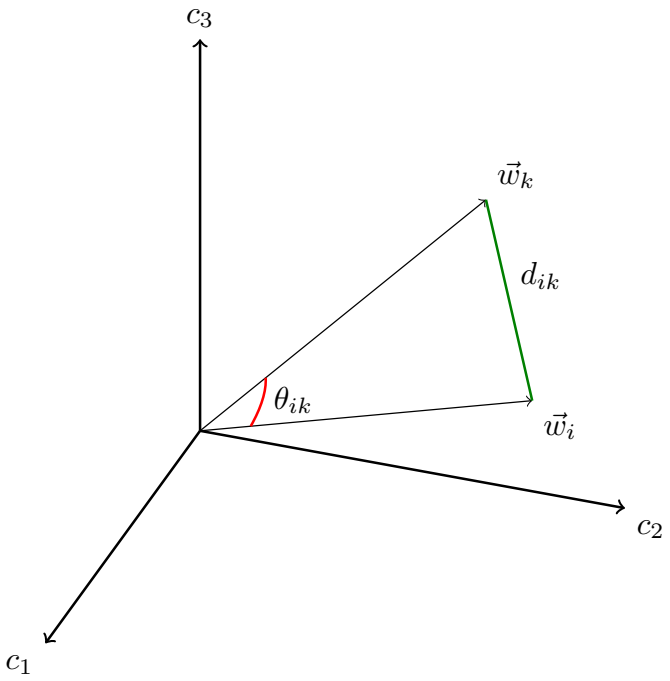


Figure 2.1.: The angle θ_{ik} and euclidean distance d_{ik} between two word vectors $\vec{w}_i$ and $\vec{w}_k$.

obtain a balanced distributional profile to determine in which contexts it can be expected to occur – i.e. its more preferred contexts – in terms of e.g. relative frequency etc. However, due to Zipf's law, the majority of words will occur very infrequently making reliable estimates difficult. When considering which words to use as context dimensions, our models are directly affected by a bias-variance tradeoff: Those words that are very frequent in a given corpus will result in richer representations (high bias, low variance). These provide overall higher similarity estimates in which interesting differences are difficult to determine. Using less frequent words as contexts will lead to very informative but less reliable estimates (low bias, high variance) (Lowe, 2001); an effect most reweighting schemes aim to ameliorate.

We can then define the context of a target word in a number of different ways. One method is to observe a naturally enclosing linguistic unit: the sentence.

Different window sizes have been shown to affect the types of relational information in the co-occurrence counts (Church and Hanks, 1990; Sahlgren, 2006),

from more syntagmatic with smaller windows to paradigmatic relations being captured better using large windows – as a higher percentage of the larger contexts are shared using larger window sizes (Lapesa, Evert, and Schulte im Walde, 2014).

Limitations of BOW Models. One effect of using smaller contexts is that they show more relational or grammatical information while larger contexts show more of the overall topic (Peirsman, Heylen, and Geeraerts, 2008). While it seems obvious that a large amount of semantic information can only be found while including the more general level of topic, the example introduced in Chapter 1 of the words *hunter* and *deer* shows that important distinctions can be obscured or even lost. We can see that while it makes sense for *hunter* and *deer* to be similar topically – as they are obviously related in the real world: we expect a deer to be acted upon by a hunter, namely by hunting, tracking or shooting it – at the same time, they clearly have a different expected relationship to one another as well as other words present in the sentences in which they would co-occur. In summary, we can say that BOW models can capture topics well, but fail to provide an adequate model of the types of comparisons and distinctions required for the modeling of selectional restriction, syntactic relations and, by extension, plausibility (e.g. of predicate-argument combinations).

Beyond BOW Models. The natural approach to address the issue of insufficient information for these tasks is to enrich or refine the context definition. Simple approaches have been to heuristically weaken the strong bag-of-words assumption and encode additional information about the observance of a word around a target. One proposal distinguishes left-context from right-context, i.e. whether the context word was observed to the left or right of a target (Patel, Bullinaria, and Levy, 1998). This could help address the *hunter/deer* issue mentioned above in a configurational language in which a standard word order reflects syntactic functions (e.g. in English, an SVO-language, subject nouns typically occur to the left of the verb whereas objects on the right).[12] A step further towards increasing contextual information is the encoding of the context word's position in relationship to

[12] A modification would be to use only left, or preceding, context (Pennington, Socher, and Manning, 2014), mimicking language models which aim to predict upcoming words.

the target (Schütze, 1993). In this family of models, the relative index would be recorded along with the context word. While Schütze (1993) showed that tasks such as part-of-speech induction can be performed using such a positional space, this definition of context incontrovertibly compounds data sparsity issues mentioned above; at the same time, the information contained in the indices themselves only implicitly model grammatical relations. Erk (2012) describes multiple possible avenues along which one could move beyond simple BOW models. Gärdenfors (2004)'s conceptual spaces are inherently vector spaces yet words are modeled into regions, and each perceptual or conceptual domain lives in its own space.

These are all grammatically uninformed models in that, syntactic relations are still latent structures that are not made use of, or explicitly captured by the models. Building on work by Grefenstette (1994a), Padó and Lapata (2007) leveraged the grammatical dependency structure for context definition and extraction, which functions to an extent as a filter on or enrichment of the contexts, similar to the positional information.

Grammatically informed models which directly encode these relations represents a more principled approach, based in linguistic theory of the structural and relational organization of language.

2.4. Structured Vector Space Models – SDSMs

As pointed out by Lowe (2001):

> *Parts of speech, and grammatical structures are also examples of latent structure in the sense that they are in principle unobservable aspects of words that reflect their distributional properties. One important direction for semantic space research is to find an appropriate type of latent structure to explain the distributional regularities that are assumed to underlie semantic similarity.*

Distributional models that explicitly encode the syntactic structure can be termed **structured DSMs** or **SDSMs**. With the inclusion of grammatical relations into the context (cf. the *hunter–deer* example given in Chapter 1), structured models are able to make more fine-grained distinctions between words that are highly related

but with important syntactico-semantic differences (Sayeed et al., 2015). This also means increasingly difficult tasks can be addressed such as estimating the plausibility of a predicate-argument combination without the need for a dedicated model.[13]

In order to obtain evidence of syntactic relations in the corpus, these latent structures must first be made explicit. This means having the corpus parsed. Dependency-based grammars are best suited to a grammatically informed context definition. The link structure of dependency edges in the parse tree provides direct access to syntactically related words in a sentence. By simply following these edges, we can obtain the bilexical relations needed to build the model. In addition, dependency parsers are very robust with recognizing predicate-argument relations which is, arguably, the basis of the sentence's semantics. Finally, dependency parsers have been developed with high accuracy and low runtime complexity (Nivre, 2008; Bohnet, 2010) making them particularly suitable for the task of parsing large corpora.

2.5. Constructing Models for Novel Languages

Our goal is to develop methods for reliably constructing SDSMs for novel languages. When moving from a linguistic resource in one language to building one in a new language a number of different approaches can be considered: **mono-** as well as **cross-lingual** methods can be explored.

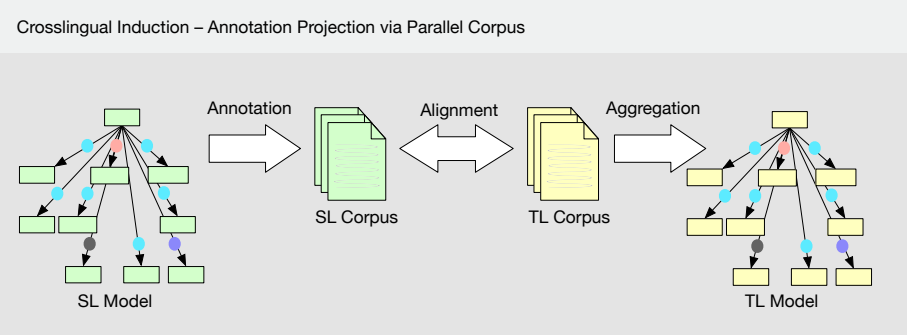


Figure 2.2.: Constructing a TL model from SL model via annotation projection

[13] In the past, such models have required head- / or relation-specific training (e.g., Resnik (1993)).

We can distinguish two types of cross-lingual transfer. If the resource consists mainly of information at the token (or instance) level – i.e. on actual occurrences of words within specific contexts – this can be achieved via **annotation projection** in parallel corpora (Yarowsky and Ngai (2001); Bentivogli and Pianta (2005)), cf. Figure 2.2. If however the resource is organized on the type (or lemma) level, bilingual dictionaries can be employed to translate the resource directly (Fung and Chen, 2004; Peirsman and Padó, 2011, , cf. Figure 2.3).

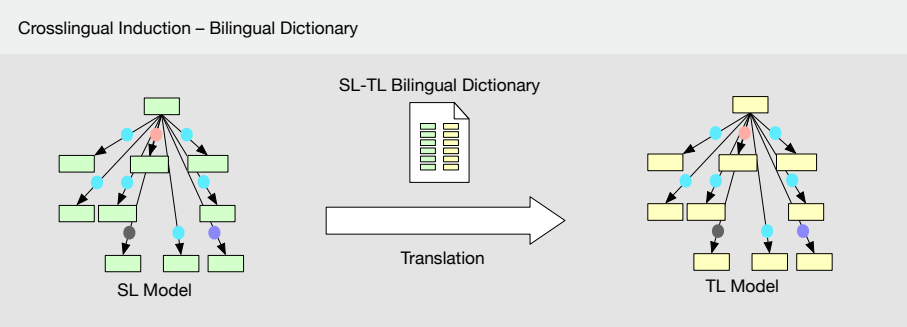


Figure 2.3.: Translation of SL model into TL using bilingual lexicon

The requirements for performing the annotation projection are quite high. Beyond the need for an aligned parallel corpus, the choice of the annotation and aggregation methods is not straightforward for a models such as SDSMs.

Another approach for building a TL SDSM is to simply implement the methodology employed in the source language to the corresponding data from the target language (cf. Figure 2.4). Obviously, in this case, neither the existing data nor the model are leveraged per se. The resource requirements on the TL side simply mirror those on the SL, but are independent of the source language – this means parallel data is no longer needed. One consequence of this fact is that performance of the target model can vary in a number of ways. Given a smaller corpus, we should expect a less robust model, i.e. lower coverage and accuracy. A TL parser with lower accuracy than in the source language would introduce noise into the TL model and lead to an additional reduction in accuracy in predictions.

The methods explored in this thesis are the monolingual and the cross-lingual translation approach using a bilingual lexicon. Finally, in addition to the monolingual (ML) and cross-lingual (XL) methods, we also explore possibilities of combining a ML model and a XL model into a multilingual model.

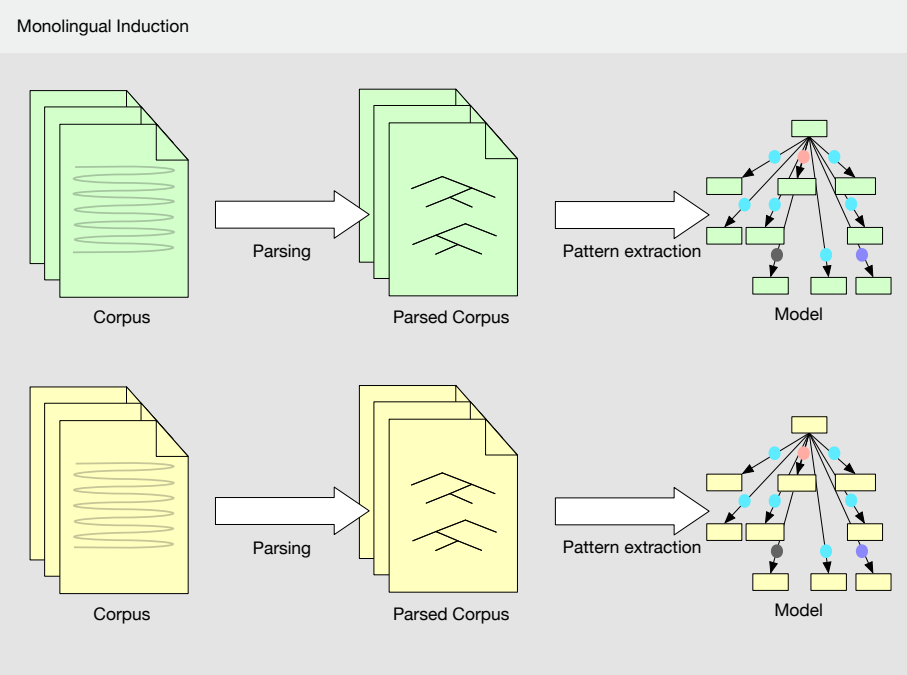


Figure 2.4.: Monolingual parallel induction of TL model

Chapter 3.

Resources

3.1. Distributional Memory as SDSM

As we have seen in the previous chapter, distributional models of word meaning represent a word's meaning as a vector of a certain dimensionality. As a result, the entire space, i.e. the set of vectors for the vocabulary at hand, takes the form of a matrix in which the numerical values indexed by the row i and column j quantify the association between the word$_i$ and context$_j$.

SDSMs such as Distributional Memory (DM) (Baroni and Lenci, 2010) then are an extension of such a two-way association by explicitly including relation labels into this association scoring. Using the terminology of Baroni and Lenci, this represents a move from a $W \times W$ or a *word-by-word* space to a W×L×W or *word-by-link-by-word* space, in which there the additional index of a **link** type. In this manner, where previously our fundamental data structure was a matrix, we now obtain a higher-order linear algebraic object, viz. a third order **tensor**.

Association scores. We define σ as the scoring function that maps a word-link-word triple to its non-negative real-valued association score:

$$\sigma : \text{W}\times\text{L}\times\text{W} \rightarrow \mathbb{R}_0^+$$

This allows for the comparison of word-link-word triples: For example:

$$\langle \textit{pencil} \text{ OBJ } \textit{sharpen} \rangle$$

is assigned a higher weight than ⟨*elephant* OBJ *sharpen*⟩.

The decisive factor in syntax-based distributional modeling is to have association scores for such triples – or in the simpler case of DSMs, word pairs – that are indicative of their true latent relationships. As discussed in Chapter 2, simple co-occurrence counts suffer the effect of highly frequent words masking the signal of less frequent but potentially more salient co-occurrences. Information theory-based association metrics such as **point-wise mutual information** (*PMI*) can be employed to counteract this effect. It measures how much information the presence of one co-occurrent conveys about the likelihood of the others, as can be seen in its definition:

$$PMI(\langle w_i \ \mathrm{L}_j \ w_k \rangle) = \log \frac{P(\langle w_i \ \mathrm{L}_j \ w_k \rangle)}{P(w_i) \cdot P(\mathrm{L}_j) \cdot P(w_k)} = \log \frac{O_{ijk}}{E_{ijk}} \begin{cases} > 0, & \text{for } O_{ijk} > E_{ijk} \\ = 0, & \text{for } O_{ijk} = E_{ijk} \\ < 0, & \text{for } O_{ijk} < E_{ijk} \end{cases} \tag{3.1}$$

where O_{ijk} represents observed frequency of the triple $\langle w_i \ \mathrm{L}_j \ w_k \rangle$in the corpus and E_{ijk} its expected frequency under the assumption of statistical independence:[1]

$$E_{ijk} = |corpus|^3 \cdot P(w_i) \cdot P(\mathrm{L}_j) \cdot P(w_k)$$

PMI achieves this by capturing the degree to which the observed frequency of the word-link-word combination differs from the expected frequency. Thus the actual frequencies of the co-occurrents are factored out. However, this also has the effect of giving too much weight to infrequent events. This can easily be seen in Equation 3.1. Using N and D to signify numerator and denominator in the definition, respectively we can see this effect in this reformulation: $PMI = \log \frac{N}{D} =$

[1]Statistical independence means each random variable varies independent from any other, i.e., $P(\langle w_i \ \mathrm{L}_j \ w_k \rangle) = P(w_i) \cdot P(\mathrm{L}_j) \cdot P(w_k)$. Equation 3.1 equates $\log \frac{O_{ijk}}{E_{ijk}}$ and $\log \frac{P(\langle w_i \ \mathrm{L}_j \ w_k \rangle)}{P(w_i) \cdot P(\mathrm{L}_j) \cdot P(w_k)}$, and by extension, due to the monotonicity of the log function: $\frac{O_{ijk}}{E_{ijk}}$ and $\frac{P(\langle w_i \ \mathrm{L}_j \ w_k \rangle)}{P(w_i) \cdot P(\mathrm{L}_j) \cdot P(w_k)}$. It is important to note that these are equivalences of *ratios* in which the corpus frequency factors present in O_{ijk} and E_{ijk} cancel each other out.

$\log(N) - \log(D)$. Let us consider the numerator $\log(N)$ as well as two probabilities in the denominator fixed. If we then significantly decrease the third probability, without loss of generality e.g. $P = P(w_i)$, then *PMI* increases with $-\log(P)$ in $\log(N) - \log(D) - \log(P)$; which in a large corpus can quickly yield very large values.[2] As indicated in Equation 3.1, positive *PMI* values correspond to triples that occur more frequent than by chance and are thus indicative of a significant relationship between the words and link.

German DM $\mathbf{DM}_{\text{DE}}$. The weighting in our German DM is informed by the definitions for the English LexDM and DepDM, using a restriction of **Local Mutual Information** (*LMI*: Evert, 2005) to $\mathbb{R}^+$. The following equation defines the estimated value of *LMI* of a word-link-word triple $\langle w_i \ \text{L}_j \ w_k \rangle$ via its observed frequency in the corpus O_{ijk} and its expected frequency E_{ijk}:

$$LMI(\langle w_i \ \text{L}_j \ w_k \rangle) = O_{ijk} \cdot \log \frac{O_{ijk}}{E_{ijk}}.$$

LMI addresses the over-estimation of infrequent events by multiplying *PMI* by the observed frequency – and can thus be described as a reweighting of the original frequencies.[3] The Distributional Memory tensor can be visualized as in Figure 3.1.

Another method of conceptualizing the DM structure is as a labeled and weighted graph, such as in Figure 3.2 which shows the eight highest-scoring contexts for the verb *push* together with their weights. Figure 3.2 shows arguments or adjuncts of $push_v$ – here, seven of eight co-occurrents occur as objects, and two as prepositional adjuncts. This graph formalization shows how DM also allows the modeling of **inverse relations**, by simply reversing edges, i.e. the edge $\langle push \ \text{OBJ} \ direction \rangle$ leads to an inverse relation edge $\langle direction \ \text{OBJ}^{-1} \ push \rangle$.

We will make use of both the tensor and graph conceptualizations of DM in this

[2] Consider the extreme case of a hapax legomenon in which it becomes perfectly associated with its co-occurrents, thus leading to a very large *PMI* value in a large corpus.

[3] *LMI* can also be seen as an approximation of the log-likelihood ratio – which is highly effective at modeling sparse co-occurrences (Dunning, 1993; Lowe, 2001) – without incurring the cost of using a full contingency table.

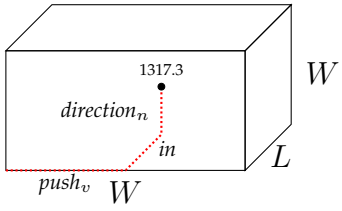


Figure 3.1.: DM as a tensor. The association scores are accessed by three indices, the word-link-word triple. (Words – not links – are subscripted by their parts of speech, noun: n, verb: v, adjective: j).

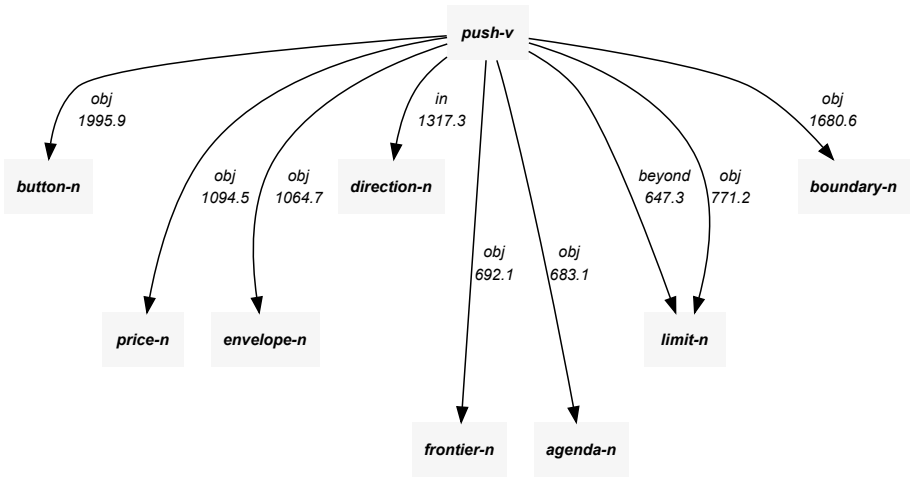


Figure 3.2.: The DepDM tensor as a labeled graph: *push* with its nine highest-scoring co-occurrents

thesis.

Tensor and matrix spaces. The benefit of such a higher-order tensor such as DM is that the model provides us with multiple vector spaces, suiting the demands of a number of distinct tasks. It follows from the earlier observation of the duality of the term-document matrix, that the three-way associations present in such an SDSM allow us to efficiently generate for a given task a tailored semantic space via a process of **matricization**.

- For example, the *word by link-word* space ($W \times LW$) contains vectors for words w_1 and its dimensions are labeled with pairs $\langle l\ w_2 \rangle$ of a link and a context word:[4]

$$\overrightarrow{w_1} = \langle\, \langle l\ w_2 \rangle : \sigma(\langle w_1\ \mathrm{L}\ w_2 \rangle)\, \rangle \tag{3.2}$$

  This space models similarity among words, and can be used in many tasks e.g. for thesaurus construction (Lin, 1998).

- Another matricized space is the *word-word by link* space (or: $WW \times L$):

$$\overrightarrow{\langle w_1\ w_2 \rangle} = \langle\, l : \sigma(\langle w_1\ \mathrm{L}\ w_2 \rangle)\, \rangle$$

  In such a space, the items of interest are word pairs $\langle w_1, w_2 \rangle$ with dimensions that are links. This space can be used to model semantic relations.

This makes the tensor structure more widely applicable in the many areas of natural language processing.

[4] Due to the fact that there is no canonical order or enumeration of the dimensions in our spaces, we opt for an index-based notation in which for a vector $\vec{v}$ whose components $x_i \in \mathbb{R}$ for a specific dimension d_i is written as $\vec{v} = \langle\, d_i : x_i \,\rangle$. E.g. the standard basis vectors $\vec{e_1}$, $\vec{e_2}$, $\vec{e_3}$ of $\mathbb{R}^3$ can be written using this notation via using the characteristic function $\chi(b) = \begin{cases} 1, & \text{if } b \text{ is true} \\ 0, & \text{if } b \text{ is false} \end{cases}$, simply as: $\vec{e_i} = \langle\, x_j : \chi(i = j)\, \rangle$, e.g. $\vec{e_2} = \langle\, x_1 : 0 \;\; x_2 : 1 \;\; x_3 : 0 \,\rangle$.The value of a component, or even its dimension, can thus easily be written as the return value of any function.

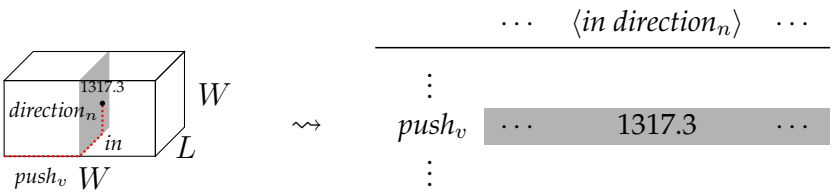


Figure 3.3.: The W×LW matricization leads to a collapse of the matrix accessed via the first word, e.g. $push_v$, into a vector indexed by single link-word pairs.

DM does not presuppose a particular vocabulary of relation type, thus leaving open the development of distinct models covering different types of relations.

Effect of link definitions. The increased informativeness of a cell in an SDSM tensor over that in a DSM matrix means the tensor will be considerably sparser, i.e., most word-link-word combinations will have a score of 0. The percentage of non-zero combinations across word and link types in the tensor is its **density**:

$$density(DM) = \frac{|\{\, \langle w_1 \text{ L } w_2 \rangle \in DM \mid \sigma(\langle w_1 \text{ L } w_2 \rangle) > 0 \,\}|}{|W| \cdot |L| \cdot |W|}$$

Taking the vocabulary as fixed, Baroni and Lenci (2010) introduced three different design choices regarding link types in the original formulation of DM.

- **DepDM** which uses dependency paths as relation types qua semantic relations (Grefenstette (1994b); Curran and Moens (2002); Padó and Lapata (2007); Rothenhäusler and Schütze (2009)). This results in approximately 800 link types and a tensor density of 0.0149%.
- **LexDM** encodes lexical relations such as hypernymy using surface patterns (Hearst, 1992) in which the participating words are directly embedded in the links. This yields over 3M link types and the resulting tensor has a density of 0.00001%.
- Finally, **TypeDM** contains links derived from the lexicalized patterns of

LexDM: explicit lexical *token*-level information is removed and the number of different token *realizations* forms the basis of the *LMI* score calculation. This significantly reduces the number of links to 25K. The resulting tensor density of 0.0005% is between the other two.

Baroni and Lenci (2010) evaluate each model type on a wide array of semantic tasks and show that DMs reliably perform comparably to or better than specialized models. Despite variation across tasks, all models perform similarly well and thus are good candidates for a general-purpose SDSM. For this reason, in choosing the basis for our TL DM we opt for the simplest definition of DM, DepDM which has the lowest barrier to implementation and also the highest density (a factor of 30 higher than TypeDM which is itself 50 times more dense than LexDM).

3.2. Dependency-parsed German corpus – SDEWAC

For the monolingual induction of a German DM, we require a large parsed corpus.

The corpus from which we extract the co-occurrence counts is the SDEWAC web corpus (Faaß and Eckart, 2013). SDEWAC is derived from the DEWAC corpus (Baroni and Kilgarriff, 2006), a large corpus collected from the German web under the `.de` top-level domain. The corpus was obtained by applying various preprocessing steps to DEWAC web corpus (Baroni, Bernardini, Ferraresi, and Zanchetta, 2008). Despite the multiple advantages of using web sites as a source for text data, a significant amount of the text thus retrieved is actually non-content data or metadata, such as copyright footers, navigation bar labels or text that is not rendered by the browser or meant to be read by a human reader.[5]

Obtaining a clean corpus. For SDEWAC, duplicate sentences originating from the same URL were removed.[6] This leads to 9 million different word types and

[5] Another example of this might be lists of keywords the creator of the site included in the source of the page intending to increase the site's ranking with search engines for certain queries.

[6] In the construction of SDEWAC sentence identification was employed to distinguish text that contains normal language from such non-content data, additionally using the criterion of being *parseable* with a high enough confidence value (Faaß and Eckart,

Rel.	Freq.	%	Rel.	Freq.	%	Rel.	Freq.	%
NK	260M	29.6	NG	5.9M	0.7	SBP	1.5M	0.17
–	160M	18.0	PM	5.6M	0.6	AC	960K	0.11
MO	120M	13.8	OP	5.3M	0.6	CVC	660K	0.07
SB	66M	7.5	DA	5.3M	0.6	NMC	570K	0.06
CJ	40M	4.6	SVP	4.4M	0.5	AMS	450K	0.05
OA	34M	3.9	PG	3.8M	0.4	DM	170K	$2 \cdot 10^{-2}$
OC	33M	3.8	RE	3.4M	0.4	PH	170K	$1.9 \cdot 10^{-2}$
CD	27M	3.1	PAR	3.2M	0.4	RS	150K	$1.7 \cdot 10^{-2}$
MNR	23M	2.6	APP	3.2M	0.4	OG	110K	$1.2 \cdot 10^{-2}$
AG	22M	2.5	CM	2.5M	0.3	VO	78K	$8.8 \cdot 10^{-3}$
PD	12M	1.4	CC	2.3M	0.3	AVC	46K	$5.2 \cdot 10^{-3}$
CP	10M	1.1	JU	2.0M	0.2	ADC	12K	$1.4 \cdot 10^{-3}$
PNC	8.3M	0.9	EP	1.9M	0.2	OA2	12K	$1.4 \cdot 10^{-3}$
RC	7.5M	0.8	UC	1.8M	0.2	SP	2K	$2.1 \cdot 10^{-4}$

Table 3.1.: TIGER (Brants et al., 2004) format dependency relation frequencies and percentages in dependency-parsed SDEWAC.

over approximately 45 million sentences (885 million tokens) parsed with the MATE German dependency parser (Bohnet, 2010). The parser was trained to optimize dependent attachment accuracy. The result of this filtering is a significant reduction in size from the unfiltered DEWAC (0.8 billion tokens in SDEWAC compared to 1.7 billion in DEWAC) but at the same time, a marked increase in the quality of the text.

Table 3.1 shows the relations present in the parsed German corpus. Approximately 20% of relation instances are subject, object and conjunction relations. The dummy relation '–' links punctuation such as commas, periods and parentheses to the preceding token; there are ca. 3.6 punctuation marks per sentence. While this significantly increases the quality of the corpus, there still are some issues such as spelling errors or non-words being present, but straining out these was not our goal.

2013).

POS	%	Explanation	Most frequent lemmas in SdeWaC
NN	20.4	common noun	*Jahr, Mensch, Zeit, Kind, Land*
ART	10.3	(in-)definite article	*der, ein, –, Der, »*
APPR	7.9	preposition; left circumposition	*in, von, mit, für, auf*
ADJA	6.0	attributive adjective	*neu, anderer, groß, erster, deutsch*
$.	5.4	sentence-final punctuation	*–, [, «, »,]*
$,	5.3	comma	*–, », [, «, o*
ADV	5.2	adverb	*auch, so, nur, noch, dann*
VVFIN	3.9	finite verb	*geben, kommen, gehen, stehen, finden*
NE	3.5	proper noun	*–, Deutschland, Berlin, USA, Europa*
KON	3.4	coordinating conjunction	*und, oder, aber, sondern, sowie*
VAFIN	3.1	finite verb	*sein, werden, haben, muss, hab*
PPER	2.7	(non-reflexive) personal pronoun	*es, sie, ich, er, wir*
ADJD	2.5	adverbial / predicative adjective	*gut, möglich, spät, wirklich, schnell*
$(	2.3	other intra-sentential punctuation	*–, «, »,], [*
VVPP	1.9	participial verb	*machen, geben, stellen, sehen, sagen*
VVINF	1.8	infinitive verb	*machen, lassen, geben, tun, sehen*
CARD	1.6	cardinal number	*zwei, 1, 2, drei, 3*
APPRART	1.5	preposition with article	*in, zu, an, von, bei*
KOUS	1.0	subordinate conjunction	*dass, wenn, daß, weil, ob*
VMFIN	1.0	finite modal verb	*können, sollen, wollen, müssen, dürfen*

Table 3.2.: Top 20 most frequent parts of speech in SdeWaC.

3.3. English-German bilingual dictionary

Following the discussion in Section 2.5, we will adopt the translational approach to cross-lingual DM construction. Translation lexicons are among the most common bilingual resources with many large ones available online in the form of crowd-sourced databases. Even in the case of unusual language pairings, it has been shown that highly accurate and large-coverage dictionaries can be obtained for virtually any language pair using probabilistic inference (Haghighi, Liang, Berg-Kirkpatrick, and Klein, 2008; Soderland, Etzioni, Weld, Skinner, and Bilmes, 2009); even in the absence of large, cleaned corpora.

The website dict.cc[7] provides numerous such lexicons, e.g. for German and English that we will use in this work. Regarding the information contained in dict.cc, we make structural assumptions about the entries when extracting translation pairs[8] resulting in a database covering 215K noun, 74K adjective and 25K verb English-German translation pairs.

3.4. Evaluation tasks and datasets

In order to determine the quality of the models, we must be able to evaluate their ability to predict linguistic realities. These can be grounded in queries at different levels of linguistic structure starting at the lexical level (e.g. obtaining semantic relatedness judgments), issues regarding linguistic processing difficulty (e.g. measuring reading times) or comparing against models of the plausibility of syntactic compositions (e.g. by mining and validating high- and low-fit arguments from corpora). These different methods of evaluation take place on different levels of syntactic complexity, as illustrated in Figure 3.5.

[7] `https://www.dict.cc`

[8] We see this as a cleaning step in which we strip such as elements as the annotations of colloquial, archaic words or removing multiword expressions and unwanted parts of speech from the line containing the translation, e.g.:

```
(alter) Gaul m jade [archaic] noun  →   Gaul-n jade-n
```

Also, the identity of a SL word's part of speech is maintained for the TL translation, e.g. we do not allow nouns to be translated into adjectives.

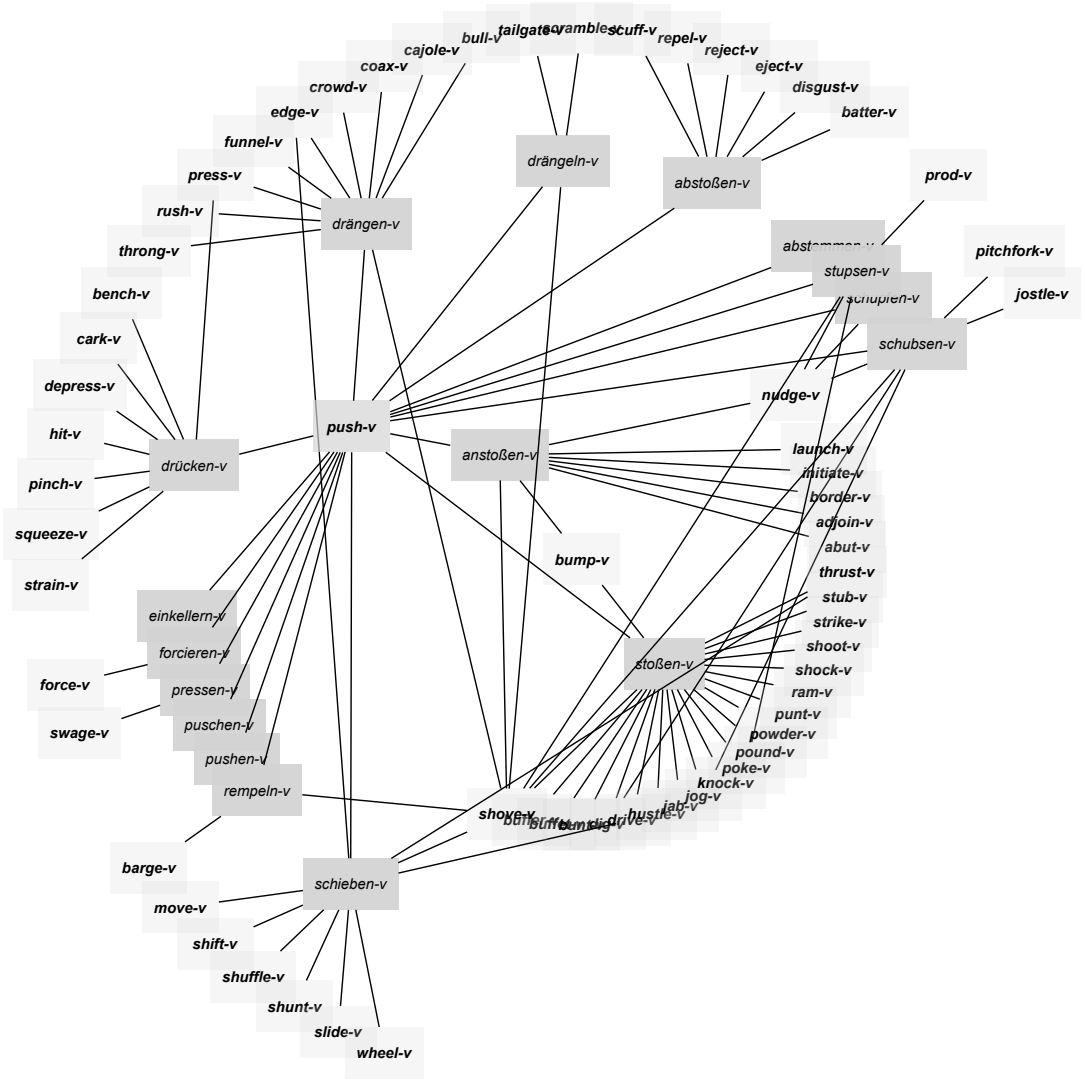


Figure 3.4.: Neighborhood of $push_v$ in dict.cc obtained by translating its German neighbors back into English. We can see that even after these two translation steps, the English words thus reached are closely related to $push_v$. At the same time, this forward and back translation multiplies into 63 new English words.

	Counts			Transl. variance			Weighted variance		
Word class	Pairs	EN	DE	bi-dir.	EN	DE	bi-dir.	EN	DE
Noun	214948	100711	130871	1.86	2.13	1.64	8.66	9.01	5.91
Verb	74448	8193	9369	2.90	3.11	2.72	7.89	7.93	6.48
Adjective	25505	37460	36260	2.02	1.99	2.05	7.00	6.45	9.16
Total	314901	146364	176500	1.95	2.15	1.78	8.15	8.27	7.07

Table 3.3.: Translations and translational variance in dict.cc. Variance is measured as number of translations for a word. Weighted variance includes corpus frequencies into a weighted average of translational variance showing that more frequent words are more ambiguous.

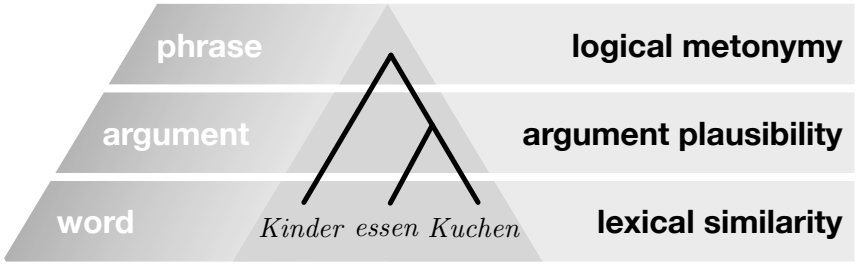


Figure 3.5.: The complexity of semantic (composition) tasks increase at higher levels of syntactic analysis. (English translation of *Kinder essen Kuchen*: kids eat cake.)

We test our models on each of these levels and in discussing the results, highlight the shortcomings and potential for the approaches investigated in this thesis.

3.4.1. Lexical relatedness

The adequacy of the lexical semantic information in (S)DSMs which aim to capture the meaning of words can be tested in a number of ways. One straightforward means of ascertaining their accuracy is to do a pairwise comparison of single word representations and compute their **relatedness**. Then any aggregate measure such as correlation coefficients, e.g. Pearson's r or Spearman's ρ can serve as a quality indicator of these estimates in testing how well they align with human judgments.

Data. For German, the Gur350 wordsim dataset (Gurevych, 2005; Zesch, Gurevych, and Mühlhäuser, 2007) is the standard dataset for testing word relatedness in German.[9] It was obtained by querying 8 human participants to judge the semantic relatedness of pairs of words on a five-point Likert scale (0 unrelated to 4 highly related). Table 3.4 gives examples of the pairs and their mean ratings and standard deviations between the ratings given for that pair. Standard deviations of 0 are present for:

- maximum relatedness scores of 4 given to synonyms:

  $Witz_n$ – $Joke_n$ $\quad$ $joke_n$ – gag_n
  $Ding_n$ – $Gegenstand_n$, $\quad$ $thing_n$ – $object_n$,

  or in cases of word derivation:

  $Erfolg_n$ – $erfolgreich_j$ $\quad$ $success_n$ – $successful_j$
  $Demut_n$ – $demütig_j$; $\quad$ $humility_n$ – $humble_j$;

- or minimum scores for random unrelated pairs:

  $italienisch_j$ – $vergehen_v$ $\quad$ $Italian_j$ – $decay_v$
  $Zebra_n$ – $Stellenanzeige_n$ $\quad$ $zebra_n$ – $job\ offer_n$
  $Kaffeetasse_n$ – $parallel_j$ $\quad$ $coffee\ cup_n$ – $parallel_j$

Interestingly, we see that even for cross-part-of-speech pairings, relatedness judgments can be very high. Figure 3.6 shows more closely how the ratings relate to the raters' agreement (evidenced by lower standard deviations between ratings) and the pairing of different parts of speech. While no significant differences in either rating or uncertainty exists across parts of speech, a general trend can be seen in which the closer the mean rating is to the extremes – i.e. either 0 or 4 – the higher the agreement.[10]

[9] Dataset is available at `https://www.informatik.tu-darmstadt.de/ukp/research_6/data/semantic_relatedness/german_relatedness_datasets/`.

[10] One might consider it desirable to select the word pairs for which the agreement was higher. In that case however, we would be excluding many of the data points and would reduce the dataset to only the extremely (dis-)similar pairs.

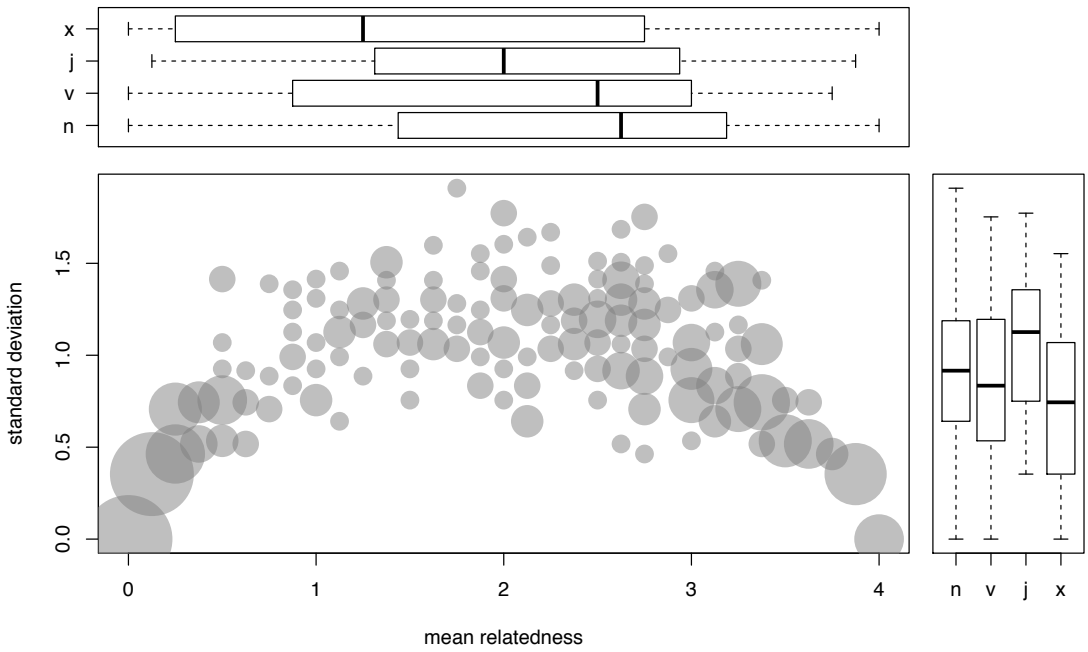


Figure 3.6.: Gur350: Rater agreement across ratings and parts of speech (average mean rating: 1.97, average rating standard deviation: .85). A lower standard deviation between ratings means a higher rate of agreement between annotators. A circle's area is proportional to the number of $\langle x, y \rangle = \langle \text{mean relatedness}, \text{standard deviation} \rangle$ pairs at that point. n, v, j and x refer to noun, verb, adjective and 'cross-part of speech' pairs, respectively.

Task and evaluation. Our model predicts relatedness as similarity scores in the W×LW vector space, i.e. cosines of the angles between words, cf. Chapter 2. The values thus obtained which range from 0 to 1 for positive-valued vectors are then correlated with the human judgments. The authors report an inter-annotator agreement of 0.69 which was measured as Pearson's r correlation can thus be taken as a theoretical upper limit for any model.

$Word_1$	$Word_2$	Mean relatedness rating	Std. deviation of ratings
deutsch$_j$ German	*Bundesbürger*$_n$ German citizen	3.5	0.756
Frage$_n$ question	*Antwort*$_n$ answer	3.25	1.389
Tierpark$_n$ animal zoo	*Giraffe*$_n$ giraffe	3	1.069
Bild$_n$ picture	*visuell*$_j$ visual	3	1.069
nachgehen$_v$ follow up	*untersuchen*$_v$ examine	2.75	1.165
Designer$_n$ designer	*Eleganz*$_n$ elegance	2.625	1.061
Design$_n$ design	*Optik*$_n$ optics	2.625	1.188
erklären$_v$ explain	*begründen*$_v$ base	2.5	1.069
Franzose$_n$ Frenchman	*deutsch*$_j$ German	2.375	1.302
Mann$_n$ man	*Geschäftspartner*$_n$ business partner	1.5	0.926
schließen$_v$ close	*Überlegung*$_v$ idea	0.875	1.126
sportlich$_j$ athletic	*teuer*$_j$ expensive	0.375	0.744
versäumen$_v$ miss	*überprüfen*$_v$ check	0.125	0.354
summieren$_v$ sum	*selbstbewußt*$_j$ self-confident	0.125	0.354
untersuchen$_v$ examine	*Benedikt*$_n$ Benedikt	0	0

Table 3.4.: Examples from the Gur350 dataset which covers 466 words – 291 nouns, 103 verbs and 72 adjectives – in 350 pairs. Pairings occur not only within a part of speech but also across.

3.4.2. Synonym detection

The task of synonym detection consists of determining from among a set of candidates which is most closely related to or synonymous with the base word.

Data. The German Reader's Digest Word Power dataset (Wallace and Wallace, 2005)[11] is a word choice dataset with 984 items. It was originally designed to help enhance readers' vocabulary in the form of a quiz: For each base word which is sufficiently infrequent in everyday parlance, the correct candidate was to be selected from among a set of superficially plausible distractors (cf. Table 3.5). This dataset is amenable to being used as a test suite for lexical semantic models, by measuring the accuracy of a model's ability to determine the target candidate, cf. Equations 3.3 & 3.4. It is comparable to the much smaller 80-item synonym portion of the English TOEFL (Landauer and Dumais, 1997b), but also contains short phrases among the candidates.

Task and evaluation. The task can be seen as an assignment task, i.e., our model aims to determine which candidate is the target. This means for every item, we can simply check whether the assignment was correct. The proportion of correct assignments is the model $\mathcal{M}$'s **accuracy** on the dataset:

$$\mathcal{D} = \{\, (w_i, \Gamma_i) \} \subset W \times \Gamma$$

the dataset with $\Gamma_i = \langle c_i^1\ c_i^2\ c_i^3\ c_i^4 \rangle$ the candidate list for w_i.

$$acc_a(\mathcal{M}, \mathcal{D}) = \frac{|\{\, d \in \mathcal{D} \mid \alpha_{\mathcal{M}}(d) = t(d) \,\}|}{|\mathcal{D}|} \tag{3.3}$$

where $\alpha_{\mathcal{M}} : \mathcal{D} \to \Gamma$ is the assignment function of the model, and $t : \mathcal{D} \to \Gamma$ identifies the true target.

If, however, our model cannot determine a unique best candidate, we have ties. In this case, we can instead frame this as a scoring task and evaluate by discounting

[11] Obtained from `https://www.informatik.tu-darmstadt.de/ukp/research_6/data/semantic_relatedness/german_word_choice_problems/`.

ties:

$$acc_s(\mathcal{M}, \mathcal{D}) = \frac{\sum_{d \in \mathcal{D}} \zeta_{\mathcal{M}}(d)}{|\mathcal{D}|} \tag{3.4}$$

where $\zeta_{\mathcal{M}} : W \times \Gamma \to \mathcal{R}$ is the scoring function:

$$\zeta_{\mathcal{M}}(w_i, \Gamma_i) = \begin{cases} |\{\, c_i^j \mid \zeta_{\mathcal{M}}(c^j) = \zeta_{\mathcal{M}}(t(d)) \,\}|^{-1}, & \text{if} \arg\max_{c_i^j \in \Gamma_i} \zeta_{\mathcal{M}}(c^j) = t(d) \\ 0, & \text{otherwise} \end{cases}$$

In the case of the scoring formulation, we can have, e.g., a three-way tie: If the actual target scored the maximal $\zeta_{\mathcal{M}}$ for the item d, Equation 3.4 results in a score of $\zeta_{\mathcal{M}}(d) = 1/3 = 0.\dot{3}$. This can be simplified to $acc_s(\mathcal{M}, \mathcal{D}) = |\mathcal{D}|^{-1}(A + \frac{1}{2} \cdot B + \frac{1}{3} \cdot C + \frac{1}{4} \cdot D)$, where A is the number of correctly predicted items with no ties, B those with a 2-way tie, C with a 3-way tie, and D with a 4-way tie. Thus acc_a becomes identical to acc_s exactly when there are no ties. We use this more general scoring accuracy measure following Mohammad, Gurevych, Hirst, and Zesch (2007).

3.4.3. Argument plausibility

Predicate-argument combinations form the basis of the structure of a sentence, e.g.

subject: *vase*$_n$–*break*$_v$

object: *eat*$_v$–*cake*$_n$

They express functional relationships between words that can – as a precursor to semantic role labeling – enable further deep processing, e.g. textual entailment and information extraction. For given predicate-argument pairs the determination can be made as to the **plausibility** of that juxtaposition.

Data. Brockmann and Lapata (2003) compiled a dataset covering three types of predicate-argument relation types: subject, object and prepositional complement.

Base word	Target	Distractor candidates		
Agility dog agility	*Hundesport* dog sport	*Aufwärmgymnastik* warm-up	*Sackhüpfen* sack race	*Formationstanz* formation dance
anachronistisch anachronistic	*zeitwidrig* contradicting in time	*sonderbar* strange	*originell* original	*zeitbedingt* temporary
clever clever	*klug gewitzt* cunning	*gewissenhaft* conscientious	*wohlhabend* affluent	*verlogen* mendacious
Cross-over cross-over	*Stilmischung* mixture of styles	*Zebrastreifen* crosswalk	*Schifferknoten* knot	*Potpourri* potpourri
Handikap handicap	*Benachteiligung* disadvantage	*Mobiltelefon-Etui* cell phone case	*Einschaltverbot* prohibition of turning on	*Störanfälligkeit* susceptibility
kultig iconic	*modern und in* modern and in	*religiös bedingt* religious	*altgewohnt* familiar	*feierlich* ceremonial
partizipieren participate	*teilhaben* take part	*Partei nehmen* advocate	*aufteilen* split up	*Brüche durcheinander teilen* divide fractions
resolut resolved	*tatkräftig* energetic	*zurückhaltend* timid	*bärbeißig* gruff	*berechnend* calculating
Server server	*Zentralcomputer* central computer	*Stromquelle* power source	*Nothelfer* first responder	*Regler* regulator
Triathlon triathlon	*Dreikampf* triathlon	*Dreier-Team* team of three	*dreitägiges Sportfest* three day athletic event	*dreieckiges Spielfeld* triangular playing field

Table 3.5.: Example items from the Reader's Digest Word Power dataset. 984 base words are paired with a set of distractors and a target synonym or gloss.

For each relation type, 10 verbs were sampled uniformly from the Süddeutsche Zeitung corpus[12]. For each ⟨*verb rel*⟩ pairing, a noun complement with a corpus-estimated high-, mid-, and low-plausibility was chosen by drawing uniformly from the equal division of log-transformed co-occurrence bands for each verb.

[12]Verbs or nouns whose relative frequency were less than 1 per million on the Süddeutsche Zeitung were not considered.

This results in 90 verb-noun pairs, 30 pairs for each relation type (cf. Table 3.6). These were rated by 61 native German speakers with an inter-subject agreement

Verb	Complement	Plausibility
	Subject	
stagnieren stagnate	*Preis* price	0.24
warten wait	*Welt* world	0.10
musizieren make music	*Grundschule* elementary school	−0.03
glitzern glitter	*Ferne* distance	−0.26
schwappen swash	*Rock* skirt	−0.60
	Object	
enttäuschen disappoint	*Gast* guest	0.31
enttäuschen disappoint	*Politiker* politician	0.22
reinigen clean	*Gehweg* sidewalk	0.20
formieren form	*Widerstand* resistance	0.19
schmieden forge	*Instrument* instrument	−0.15
	PPobject	
erkranken_an contract	*Malaria* malaria	0.35
teilnehmen_an participate in	*Seminar* seminar	0.35
kommen_zu come to	*Schluß* conclusion	0.22
denken_an think about	*Kleinigkeit* detail	0.19
riechen_nach smell like	*Runde* round	−0.51

Table 3.6.: Examples from the argument plausibility dataset (Brockmann and Lapata, 2003).

(Pearson's r) of .79 for subject, .81 for object, .82 for prepositional object and .81 overall). These agreement values, again, serve as an upper limit for model performance.

Task and evaluation. As opposed to the simple lexical relatedness models of the matricized DM in the lexical tasks outlined above, the model's knowledge of relation types can be used to generate relation-specific models $\mathcal{M}_r$.

Our task is to correlate model predictions of predicate-argument combination plausibility with human judgments:

$$cor(\mathcal{M}, \mathcal{D}) = cor(\langle \mathcal{M}_{r_i}(v_i, n_i) \mid (r_i, v_i, n_i) \in \mathcal{D} \rangle, \langle judg_i \rangle)$$

3.4.4. Logical metonymy

Metonymy is the phenomenon of one word being substituted for another (Lakoff and Johnson, 1980; Panther and Radden, 1999):

We went for a drive after John showed us his new wheels. (= car)

Berlin took a position against the measure. (= the German government)

In the case of logical metonymy, the sentential construction is a shorthand for an **covert event** (CE) that is supplied by the reader but not linguistically realized at the surface level:

The author began the novel. (= e.g. writing / reading)

Americans love baseball. (= e.g. watching / playing)

Verbs that can trigger metonymic interpretations allow for a verbal complement. Common classes are psychological verbs, e.g. *enjoy, hate*, and aspectual verbs, e.g. *begin, stop*. The specific real-world event remains implicit in the construction and must be provided by the reader's understanding on the basis of the relevant linguistic material present in the sentential – or surrounding – context. The process of CE interpretation – we can be described as the triggering and recovery of the implicit event – is dependent upon the combinations of multiple contextual participants, i.e. the subject and object of the metonymic verb as well as the metonymic

	CE	
	high thematic fit	low thematic fit
Der Bergsteiger versuchte, den Berg The alpinist tried the mountain	*zu erklimmen.* to climb.	*zu malen.* to paint.
Der Künstler versuchte, den Berg The artist tried the mountain	*zu malen.* to paint.	*zu erklimmen.* to climb.

Table 3.7.: Structure of logical metonymy (Zarcone et al., 2012) dataset: Each object is matched with two subjects and two covert events, here *Berg* with *Bergsteiger / Künstler* and *erklimmen / malen* and one metonymic verb (here: *versuchen*). In such a 2x2 design, each subject occurs once in a high-typicality and in a low-typicality combination. The same holds for each of the covert event-denoting verbs.

verb itself. By being highly sensitive to its syntactic context (Zarcone and Padó, 2011; Zarcone, Padó, and Lenci, 2012), a model of logical metonymic interpretation would go beyond individual predicate-argument plausibility considerations, e.g. in the evaluation against the Brockmann and Lapata (2003) dataset. The linguistic phenomenon of logical metonymy is thus well-suited for testing syntax-aware computational models of semantics, such as SDSMs, that can potentially help shed light on the cognitive processes involved in semantic interpretation.

Data. The dataset consists of 96 $\langle s, v, o, e\rangle$ tuples. These represent 24 sets of four subject, metonymic verb, object and covert event groups. Table 3.7 shows the structure of the 2x2 design underlying the dataset. The verb arguments were produced by 20 participants given only the object and the subjects were elicited by 10 participants as the most expected agents for each given a particular $\langle o, e\rangle$ combination. Each of the 24 objects with its randomly chosen metonymic verb gives rise to two high-thematic fit $\langle s, v, o, e\rangle$ tuples and two low-thematic fit ones. **Thematic fit** is the degree of compatibility of a complement as the filler of a specific thematic role given a predicate. It thus depends on the filler, the predicate and the role as well as potential additional context information. We can reasonably approximate these thematic or semantic roles required for an assessment of thematic fit via the dependency relations already present in our

Subject	Object	Metonymic verb	Covert event	Typicality
Bergsteiger	*Berg*	*versuchen*	*erklimmen*	high
Künstler	*Berg*	*versuchen*	*erklimmen*	low
Künstler	*Berg*	*versuchen*	*malen*	high
Bergsteiger	*Berg*	*versuchen*	*malen*	low
climber / artist	mountain	try	climb / paint	
Ober	*Saft*	*aufhören*	*eingießen*	high
Baby	*Saft*	*aufhören*	*eingießen*	low
Baby	*Saft*	*aufhören*	*trinken*	high
Ober	*Saft*	*aufhören*	*trinken*	low
waiter / baby	juice	stop	pour / drink	
Handwerker	*Fenster*	*probieren*	*einbauen*	high
Hausfrau	*Fenster*	*probieren*	*einbauen*	low
Hausfrau	*Fenster*	*probieren*	*putzen*	high
Handwerker	*Fenster*	*probieren*	*putzen*	low
workman / housewife	window	try	install / clean	
Bäuerin	*Apfel*	*anfangen*	*pflücken*	high
Bäcker	*Apfel*	*anfangen*	*pflücken*	low
Bäcker	*Apfel*	*anfangen*	*schälen*	high
Bäuerin	*Apfel*	*anfangen*	*schälen*	low
farmer / baker	apple	start	pick / peel	

Table 3.8.: Examples $\langle s, o, v, e \rangle$ tuples and their typicality from German logical metonymy dataset (Zarcone et al., 2012)

SDSM, by taking e.g. the transitive subject of a verb as the agent of the denoted event, or a direct object as its patient. With this approximation, we can talk about thematic fit in terms of the **typicality** of the syntactic filler for that relation, without the need for an extra role labeling step in processing.

It was shown that the typicality assignments are reflected in self-paced reading times (Zarcone and Padó, 2011) as well as in probe recognition latencies (Zarcone et al., 2012). Typical events were associated with lower reading times and slower rejections as probe words after sentences which evoke them.

Task and evaluation. We use the **expectation composition and update** model (ECU; Lenci (2011)) which transforms the pairwise relations present in DM into a model of multiple links in the DM graph using a stepwise combination of association scores. Conceptually, ECU compares a synthetic prototype vector against the proposed slot filler. To construct the prototype, ECU first computes the verb's association scores – which are interpreted as **expectations** – for its object slot and then combines these values with the subject's expectations for the object. Formally, we arrive at the composed verb's expectations for the object as:

$$EX_V(v) = \lambda o.\sigma(\langle v \text{ OBJ } o \rangle)$$

The subject's expectations for the object are:

$$EX_S(s) = \lambda o.\sigma(\langle s \text{ VERB } o \rangle)$$

Then the updated expectation is defined as:

$$EX_{SV}(s,v) = \lambda o.\sigma(EX_V(v)(o)) \circ \sigma(EX_S(s)(o))$$

where either an additive or multiplicative combination function $\circ$.

An accurate prediction for a pair – with high- and low-typicality objects o_h and o_l – would be model expectations: $EX_{SV}(s,v)(o_h) > EX_{SV}(s,v)(o_l)$. Then the centroid, i.e. vector sum, of the k highest expected objects is defined as the prototype.

Part II.

Mono- and Cross-lingual Induction of SDSMs

Chapter 4.

Monolingual DM Induction

In this chapter, we discuss the process of designing our monolingually induced structured distributional semantic model from a German web corpus. Finally, we describe the resulting Distributional Memory, DM_{DE}, in terms of node and link frequencies as well as properties of the matricized vector spaces and compare and contrast our German DM with Baroni and Lenci (2010)'s English models.

4.1. Inducing a DM from a German corpus

Model type selection. The first step in inducing the monolingual German DM was to settle on one of the three DM variants proposed by Baroni and Lenci (2010). While the DM framework does not assume any specific source for the tuples, syntactic relational information is required. As noted in Chapter 3, the patterns in LexDM and TypeDM are more complex than those in DepDM. In the construction of the former, many manual or semi-manual annotations were used, including a list of high-frequency verbs chosen to be part of the lexicalized edge labels (Baroni and Lenci, 2010), i.e. labels containing lexical information. While the more intricately designed lexicalized models[1] nearly always proved to be the top performing DM types in their study, the significantly less complex DepDM performs at similar levels on many tasks, e.g. synonym detection (Baroni and Lenci, 2010, p. 694), noun clustering (Baroni and Lenci, 2010, p. 696) and predicting verb–argument plausibility judgments (Baroni and Lenci, 2010, p. 698). For these reasons we opted to follow a DepDM-style approach with simple, i.e. short-path,

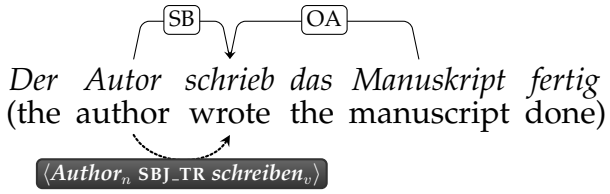
[1] These are models containing such lexicalized edge labels.

lexical link patterns for our model in addition to the grammatical patterns using the functional relations in the dependency graphs. Our objective by doing this is to make the construction and structure of DMs in novel languages analogous: The intuition being that shorter paths should transfer across language barriers better than longer ones. At the same time, this should also reduce the negative impact on model quality due to any issues that impact the TL parser's accuracy.

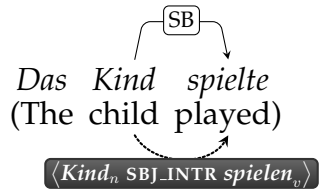
Defining patterns to extract links. The types of German links we include correspond more or less directly to the straightforward syntactic patterns described in the outline of DepDM. These patterns were obtained by collecting the most frequent syntactic structures within a large German corpus (cf. Section 3.2). The extracted patterns can be categorized into one of two groups: **unlexicalized** and **lexicalized patterns**.

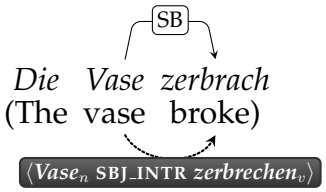
Unlexicalized patterns. We use unlexicalized patterns to determine a total of seven link types which extract the syntactic configurations both at the verb phrase and at the sentence levels, cf. (Baroni and Lenci, 2010, p. 686f.). These are:

SBJ_TR subjects for transitive verbs; usually actors, e.g.

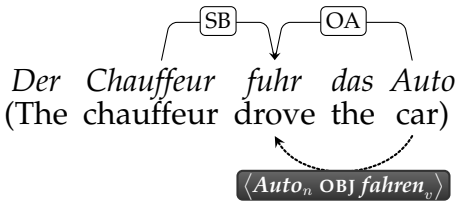


SBJ_INTR subjects for intransitive verbs; usually actors or themes, e.g.

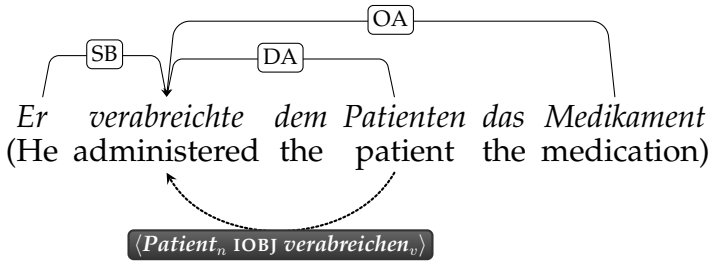




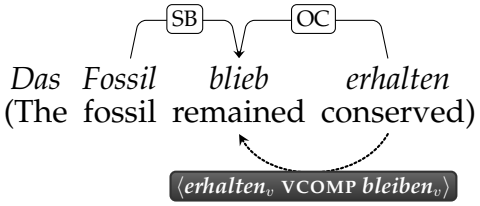
OBJ direct objects; e.g.



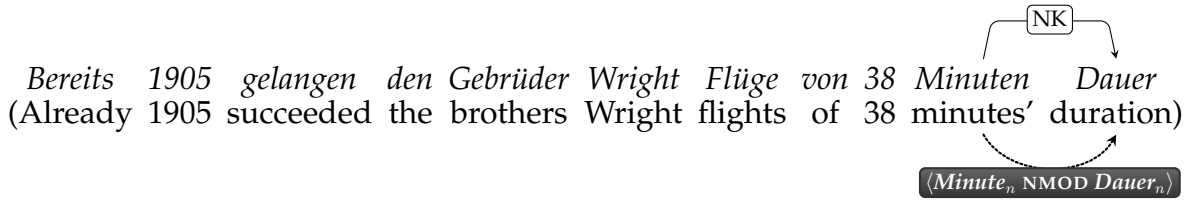
IOBJ indirect objects; e.g.

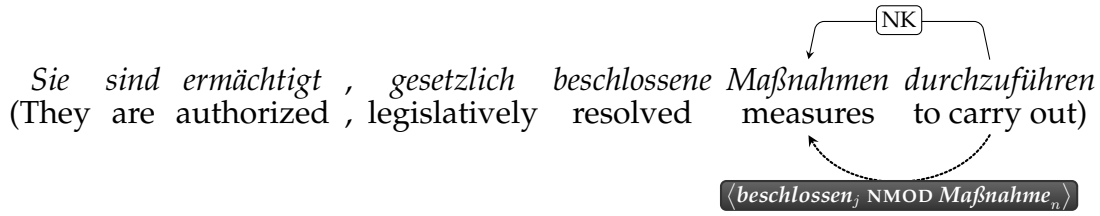


VCOMP verb complements; e.g

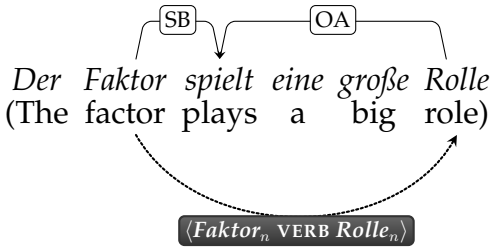


NMOD noun modification; e.g.





VERB the relation between the subject and object of a transitive verb construction; e.g.



These patterns are termed unlexicalized as their link types do not directly contain lexical information. The links are in direct one-to-one correspondence to their pattern's description of the syntactic context.

Lexicalized patterns. A lexicalized pattern can belong to one of two groups. They either:

- incorporate surface lexical information – which e.g. in patterns such as

  $$\boxed{\text{SURFACE}: \ \text{PAT}^{\text{W}_1}\ [\ldots]\ \text{PAT}^{\text{L}}\ [\ldots]\ \text{PAT}^{\text{W}_2}}$$

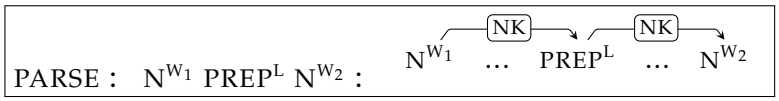
  resulting in an instantiated surface edge: $\langle w_1\ \text{L}\ w_2 \rangle$ in which each element matches its corresponding sub-pattern with optional additional patterns [...]; or

- use dependency edges to extract lexical material according to:

  $$\boxed{\text{PARSE}: \ \text{PAT}^{\text{W}_1}\ \text{PAT}^{\text{L}}\ \text{PAT}^{\text{W}_2}\ :\ \overset{\alpha}{\leftarrow}\ \text{PAT}^{\text{W}_1}\ \ldots\ \text{PAT}^{\text{L}}\ \ldots\ \overset{\beta}{\rightarrow}\ \text{PAT}^{\text{W}_2}}$$

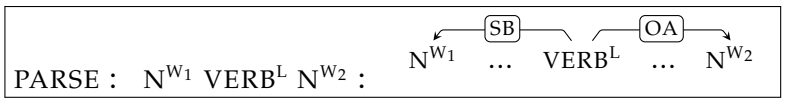
  for the declared values and orientations for the edges α and β (cf. below for examples).

The following descriptions define the three remaining lexicalized patterns. The pattern most frequently instantiated in SDEWAC is



which captures the relational chain between two nouns via a preposition. This patterns covers phrases such as *Recht auf Auskunft* (right to disclosure) which results in the triple ⟨*Recht* AUF *Auskunft*⟩ where the link type is instantiated by the lexical realization of the preposition, in this case AUF.

The second most frequent lexicalized pattern is:



Here, the link is set to be the particular verb encountered with the subject $\boxed{N_1}$ and object $\boxed{N_2}$. The sentence *Hochtief erlebt Aufwind* (Hochtief experiences upturn) for instance results in the triple ⟨*Hochtief* ERLEBEN *Aufwind*⟩.

Finally, a pattern designed to link adjectives to noun combinations is

$$\boxed{\text{SURFACE : ADJ}^{W_1}\ \text{N}^{W_2}\ \text{VON (ART|ADJ)}^*\ \text{N}^{L}}$$

which extracts the second noun involved in the combination as the triple's link.[2] It covers such phrases as *heutige Größe von dem Stern* which yields the triple ⟨*heutige* STERN *Größe*⟩.

The presence of lexical items in the links means that these patterns will yield a very large number of link types. Note, however, that we have ignored prepositional arguments of verbs, as well as adverbs and relative clauses. On the one hand, the rationale for this decision lies in the attempt to minimize the impact of problematic parser output. On the other hand, as we are only building a model for nouns, verbs and adjectives, the patterns defined above already represent the majority of the parsed corpus. The lexicalized patterns match in the case of approximately 38% of sentences in SDEWAC while the unlexicalized ones cover 98% of the sentences. At the same time, we are also motivated by the idea of minimalism: Keeping things simple and straightforward to define means a lower barrier to the implementation of similar patterns in novel languages. By restricting our models to using only these patterns, we wanted to find out whether – and did find that – a restricted

[2]The sub-pattern (ART|ADJ)* is written as a regular expression and matches a sequence of zero or more instances of articles and adjectives.

selection of highly frequent relations still enables our models to perform well.

In addition to these considerations, we also discovered that German – as opposed to English – offers a range of difficulties when extracting word relations from parsed text.

In this thesis, we opt to remain as close to the original data as possible by not attempting to simplify noun compounds or verbal complexes – thus saving time and effort on searching for the optimal preprocessing steps – while at the same time introducing the risk of an increase in data sparsity.

4.2. Qualitative Analysis of DM_{DE}

The resulting monolingually induced German DM contains over 78M links for 3.5M words (nouns, verbs and adjectives).

It contains approximately 220K link types, almost all of which stem from the lexicalized patterns, with the mean number of links being 22 per lemma. A comparison with the original English DM with 130M links covering 30K lemmas and just over 25K link types means that DM_{DE} is considerably more sparse.

	Unlexicalized links	
N (I)OBJ V	7,833,635	noun n is (in)direct object of verb v
N SBJ_(IN)TR V	7,478,550	noun n is subject of verb v in an (in)transitive construction
V_1 VCOMP V_2	677,397	subcategorization of verb v_1 by a verb v_2 (excluding modals and auxiliaries)
X NMOD N_2	1,575,516	noun n_2 modified by $x \in \{adj, n\}$
N_1 VERB N_2	3,304,045	noun n_1 is subject and n_2 object of a verb
	Lexicalized links	
link types	220,269	e.g. IN (in), VON (from), AN (to/on), ALS (as), VOR (before/in front of), UM (about/around), GEGEN (against), STELLEN (to put), MACHEN (to make/do), BIETEN (to offer), GEBEN (to give)
links	18,180,604	2,462,927 $\boxed{N_1 \text{ IN } N_2}$, 1,424,398 $\boxed{N_1 \text{ VON } N_2}$, 1,170,609 $\boxed{N_1 \text{ MIT } N_2}$, ...

Table 4.1.: Statistics for the German DM_{DE} tensor. Approximately half of the resulting links are lexicalized.

Table 4.1 lists the frequency of the lexicalized and unlexicalized link types. As described in Section 4.1, there are over 20M unlexicalized triples for our 7 link types and over 18M lexical triples for the more than 220K lexicalized link types.

SBJ_INTR^{-1}		SBJ_TR^{-1}		OBJ^{-1}	
Realität$_n$ reality	2,339.3	*Entwurf*$_n$ draft	4,455.5	*Chance*$_n$ opportunity	18,805.5
Wirklichkeit$_n$ actuality	1,343.0	*Mensch*$_n$ human	3,555.2	*Film*$_n$ film	14,978.0
Sache$_n$ thing	1,204.1	*Senat*$_n$ senate	3,244.1	*Bild*$_n$ picture	14,925.5
Welt$_n$ world	802.4	*Zuschauer*$_n$ spectator	3,147.2	*Möglichkeit*$_n$ possibility	12,429.9
IOBJ^{-1}		VCOMP		VCOMP^{-1}	
Mensch$_n$ human	512.7	*bekommen*$_v$ get	6,044.2	*konfrontieren*$_v$ confront	9,822.6
Tatsache$_n$ fact	439.3	*geben*$_v$ give	2,972.1	*aussetzen*$_v$ expose	8,822.0
Wahrheit$_n$ truth	382.7	*sagen*$_v$ say	1,939.7	*zwingen*$_v$ force	6,742.3
Zukunft$_n$ future	324.8	*glauben*$_v$ believe	1,024.5	*kommen*$_v$ come	4,061.0

Table 4.2.: Highest-*LMI* co-occurrents for the verb *sehen* (see).

Table 4.2 shows the highest scored neighbors for the verb *sehen* in DM.DE. We can see that besides the concrete subjects (e.g. *Mensch* and *Zuschauer*) and objects (e.g. *Film* and *Bild*) we also have a number of abstract nominal complements that arise from a more metaphoric use of the verb (e.g. *Chance*, *Tatsache* and *Möglichkeit*). The intransitive subjects and indirect objects are mainly due to misparses, e.g. *'die X sieht ... aus'* (The X appears/looks like ...) and idiomatic expressions, respectively, e.g. *'der Y ins Gesicht sehen'* (to face/confront Y); this is also the case for *sagen*, which is usually linked as a complement when there is a sequence of multiple verb-final clauses, or a clause with a parenthesis, in which *sehen* follows *sagen*, but not always in an actual complementizer relation e.g.:

Plato [...] ***sagt***, *daß er die Weltenseele an den Weltenleib gekreuzigt* ***sieht***, [...].

Plato [...] **says** he **sees** the world's soul crucified to the world's body, [...]

> [. . .] *Werkstatt des Britischen Informationsmmisteriums , das - ganz im Vertrauen* ***gesagt*** *- seine Aufgabe darin* ***sah*** *, das Denken möglichst vieler Menschen zu beeinflussen .*
>
> [. . .] workshop of the British Ministry of Information which - **speaking** in the strictest confidence - **saw** as its responsibility to influence the thinking of as many people as possible .

The VCOMP^{-1} examples come from phrases such as *sich mit X konfrontiert sehen* (to be confronted with X) or *sich Y ausgesetzt sehen* (to feel exposed to Y). There are thus multiple senses ranging from concrete to abstract and metaphoric among the verb's most associated neighbors.

Tensor density. One measure that can be used to characterize the resulting DM is via its density. A tensor's density indicates the percentage of its cells that contain non-zero weight and can be understood as the probability of a uniformly sampled triple being assigned a non-zero score or being present in the DM graph. Baroni and Lenci (2010) provide the densities for their three English DM variants, all of which cover approximately 30K English nouns, verbs and adjectives, varying thus most significantly in the identity and number of links types:

English DM variant	# link types	density
DepDM	796	$1.49 \cdot 10^{-4}$
TypeDM	$25,336$	$5 \cdot 10^{-6}$
LexDM	$3,352,148$	$1 \cdot 10^{-7}$

As we do not restrict the lexical content for our German DM, we have a larger number of word nodes which results in a lower overall tensor density. The density of DM_{DE} with $3,543,603 \times 442,032 \times 3,543,603 = 5.6 \cdot 10^{18}$ possible word-link-word combinations and $78,745,438$ attested tuples is: $1.41867 \cdot 10^{-11}$, which is significantly less dense than the English DMs. One obvious reason is the difference in vocabulary sizes with the unrestricted use of the German web corpus – which still contains noise on the word level – leading to a vocabulary size over 100 times larger than the English DMs.

If we subsample the 30K most frequent as our nodes, we can compare densities more directly with the original DMs. The density of DM_{DE} reduced to the top 30K

most frequent words in SDEWAC contains 306,679 link types and has a density of $1.79 \cdot 10^{-5}$ which is a higher density than TypeDM with an order of magnitude higher number of link types. Thus we can achieve a comparable level of graph connectivity to TypeDM and DepDM on the highly frequent words in the corpus but can also achieve higher coverage by including more lexical material in the link labels.

Analysis of W×LW vectors. Finally, we assess the information available on the basis of which lexical semantic predictions can be made. Among DSMs, vector similarity is one of the most common methods of obtaining predictions.

To this end, we select three frequency bands, of high, mid and low frequency ranges for our test words. Their frequencies consist of context counts, i.e. the number of contexts in the W×LW matricization that are non-zero.[3]

We select 100 words per band and calculate the pairwise similarities within each band and across all of them and report summary statistics on those values. In addition, we also provide statistics for the formal properties of the W×LW matricization of DM.DE and compare it with the English DM provided by Baroni and Lenci.[4]

Results. We find that in terms of the number of edges per word, the English DM has a approximately four times as many as DM.DE. This factor of diminished density corresponds to the tensor density values given above and is consistent across frequency bands.

With respect to the vectors' L^2 norms[5], they are much more comparable (cf. their mean or median values). Interestingly, the maximum values for DM.DE are significantly higher than in the English DM. This means that there are a small number of vectors that have a number of edges with high association scores, which

[3]These counts correspond to the number of non-zero components in a target's vector in the matricized vector space. The reason for using these values as opposed to corpus frequency is that in the case of cross-lingually constructed DMs, there is not TL corpus underlying the model.

[4]Available from `http://clic.cimec.unitn.it/dm/`.

[5]This is the standard norm corresponding to the geometric length of a vector: $\|\vec{v}\| = \sqrt{\sum_i v_i^2}$.

	high	mid	low	across
# Edges per word				
min	2415	792	72	
median	5343.00	1234.50	471.00	
mean	11986.83	1396.71	461.85	
max	98095	2327	780	
Word vector L^2 norm				
min	422.29	183.29	88.78	
median	1120.46	431.09	249.25	
mean	2452.96	504.08	253.61	
max	27512.04	1597.64	596.79	
Cosine similarity				
min	0.00	0.00	0.00	0.00
median	0.04	0.01	0.01	0.01
mean	0.06	0.02	0.01	0.03
max	0.53	0.37	0.35	0.53
non-zero sims	97.6%	90.6%	80.4%	91.0%

Table 4.3.: English TypeDM W×LW stats

	high	mid	low	across
# Edges per word				
min	673	295	1	
median	1298.50	416.00	59.50	
mean	3356.49	454.36	96.77	
max	35619	672	283	
Word vector L^2 norm				
min	202.35	171.88	11.17	
median	1266.57	287.73	131.25	
mean	4426.46	462.90	200.19	
max	95057.71	2681.19	3692.12	
Cosine similarity				
min	0.00	0.00	0.00	0.00
median	0.01	0.01	0.00	0.00
mean	0.02	0.01	0.00	0.01
max	0.84	0.35	0.17	0.84
non-zero sims	62.3%	73.0%	12.2%	43.8%

Table 4.4.: DM.DE W×LW stats on top 30K words.

we can deduce from there being fewer edges available compared to the English DM among which to distribute these high norm values.

Finally, cosine similarity values are more frequently zero in our German DM, unsurprisingly due to the lower edge counts per word. On average, similarities are a factor of approximately 3 higher in English – except for the case of the maximal similarity in the high band, markedly higher than in the English DM.

To summarize, we have shown that with simple patterns we can get a broad coverage SDSM that, when reduced to the vocabulary same size, is comparable to the medium-complexity English DM of Baroni and Lenci (2010) DM in terms of size as well as in a number of formal properties. In the next chapter, we will investigate methods for bootstrapping the available English DM into a TL DM.

Chapter 5.

Cross-lingual Induction of DM

In the last chapter, we discussed the design decisions in our approach to monolingually inducing a DM for German and discussed the resulting tensor's characteristics. To achieve this result, we required a large TL corpus, a reliable TL dependency parser as well as the definition of lexicalized and unlexicalized dependency patterns from which to build the model. The present chapter introduces our methods for significantly lowering these requirements relying instead solely on the existence of a SL DM and a bilingual dictionary.

As briefly noted in Chapter 2, we see two main types of approaches to learning TL models from SL data cross-lingually. The first is annotation projection which necessitates the existence of a (quasi-) parallel corpus. In addition, further investigations must be undertaken into methods to aggregate the projected instance-level annotations into type-level ones. On the other hand, there exists the second possibility of translating the source model directly.

Each approach makes assumptions, engendering corresponding tradeoffs, on what resources are available and whose information can have an advantageous impact on the resulting TL model. We want to make as few assumptions on resource availability as possible, as we would like to maximize the potential number of TLs to which these methods can be applied. Furthermore, while large accurately parsed, high-coverage parallel corpora currently still seem a far-off notion, there has been significant work on constructing high-quality bilingual dictionaries between arbitrary language pairs (Sadat, Yoshikawa, and Uemura, 2003; Haghighi et al., 2008; Xia, Lewis, Goodman, Slayden, Georgi, Crowgey, and Bender, 2016).

5.1. Methods

In the situation where some resource is available in a source language, but not in a target language, two strategies can be taken to create the TL resource. The first one is **parallel induction**. The schema applied in the creation of the original SL resource is reproduced, as closely as possible, for TL (recall Figure 2.4). Such an approach was taken in the last chapter.

The second one is **cross-lingual transfer** in which the existing SL resource is transferred directly into the TL. This strategy represents an attractive approach as it directly leverages the existing resource. As we have argued, translation is much more attractive an approach. By adopting English as the source language we can leverage the more favorable position of English on the resource gradient, that is, the higher level of maturity of English NLP models, methods and data, in particular of corpora and parsers, compared to most other languages.

For many languages, the treebanks required to develop and train parsers have become available only within the last decade or so if at all (Buchholz and Marsi, 2006), while English has been at the forefront of NLP development for several decades. Multiple highly accurate dependency parsers exist for English (McDonald, Pereira, Ribarov, and Hajič, 2005; Nivre, 2006, e.g.,). At the same time, distributional semantic modeling has thrived in English with arguably the widest range of large and cleaned corpora of any language.

By adopting such a translational approach, our induction methods should become applicable to as many target languages as possible, thus improving the resource particularly for those languages for which very few resources are available. The cross-lingual methods we develop here work without any target language corpora, either monolingual or bilingual. The only requirement beyond the original DM is a simple translation lexicon, i.e. a list of translation pairs and without even the need for translation probabilities.

5.1.1. Challenges for translation-based transfer

For the language pair at hand, English–German, a number of reliable, high-coverage bilingual dictionaries are publicly available that have been previously employed in NLP tasks, e.g. dict.cc.

Lang.	Word	#*tr*.	Example translations
DE	*Dummkopf*$_n$	61	*fool*$_n$, *ass*$_n$, *dummy*$_n$, *idiot*$_n$
	Trottel$_n$	59	*charlie*$_n$, *zombie*$_n$, *wally*$_n$, *fool*$_n$
	Schlag$_n$	56	*song*$_n$, *type*$_n$, *attack*$_n$, *hit*$_n$
EN	*line*$_n$	52	*Grenze*$_n$, *Richtung*$_n$, *Artikel*$_n$, *Produkt*$_n$
	branch$_n$	46	*Arm*$_n$, *Abteilung*$_n$, *Branche*$_n$, *Sektor*$_n$
	pitch$_n$	45	*Platz*$_n$, *Höhe*$_n$, *Feld*$_n$, *Grad*$_n$

Table 5.1.: Words with the highest degree of translational variance (#*tr*.) in the EN-DE portion of `dict.cc` ranked by corpus frequencies (DE: SDEWAC, EN: WACKY (Baroni et al., 2008)).

The main issue within this approach is that simple word-based translation runs into serious ambiguity problems. Translating a DM model would mean transforming a SL graph – consisting of the weighted (un-)lexicalized edges and nodes containing the SL lemmas – into a TL graph such that the syntactic and semantic relationships between them remain intact. In cases where these translations are not just synonyms but in fact express distinct senses of the English word (e.g. *wood*$_n$ as material: *Holz*$_n$, as an area: *Wald*$_n$), the English node would need to be 'split', with all its incoming and outgoing edges assigned to either of the two German nodes. At first glance, this is equivalent to a full-fledged sense disambiguation problem making it completely non-trivial to solve. Certain types of lexical information would be of use in such a task, such as sense distinctions or common collocations. Having translation pairs grouped by sense would allow a restriction of the translation multiplication to remain strictly within a sense. Beyond involving a significant amount of effort and fine-tuning of the methods, the translation of collocates could function as the basis for filtering the edges on the TL side but would significantly change the structure of the graph by splitting and merging subgraphs.

We will work with the simplifying assumption that syntactic relations show a high degree of parallelism for the language pair English–German in particular, but also in general, following the **direct correspondence assumption** introduced by Hwa, Resnik, Weinberg, Cabezas, and Kolak (2005).[1]

[1]Their work dealt with aligned translated sentence pairs and found it reason-

This is unproblematic for one-to-one translations where nodes are just relabeled, e.g. $wrist_n \rightarrow Handgelenk_n$. However, as soon as the one-to-one correspondence breaks down, the need for resolution strategies becomes evident. For instance, when two English lemmas correspond to a single German lemma – $\{ arm_n, branch_n \} \rightarrow Arm_n$ – the two English nodes must be collapsed in some manner. At the same time, 38% of the English words in dict.cc have more than one translation in German, and 29% vice versa.

Figure 5.1 shows the translational variance, i.e. the number of words a word translates to. As with word frequency, we find also here a Zipfian distribution, i.e. in both languages the rank order of the number of translations comports approximately according to an inverse power law with those translational variance values.

On average, an English lemma has 2.3 German translations, and the average German lemma has 1.9 English ones. Corpus frequency correlates moderately with translational variance in both German (Spearman's ρ: .46, $p < .001$) and English (Spearman's ρ: .49, $p < .001$).

A look at the words from dict.cc presented in Table 5.1 reveals an interesting phenomenon, namely that measured translational variance in the resource is not necessarily indicative of true underlying semantic ambiguity. Due to the fact that the lexicon is constructed as a crowd-sourced project we find that beyond highly ambiguous words such as $line_n$ (EN #1) and $branch_n$ (EN #2) or $Schlag_n$ (DE #3) and $Verbindung_n$ (DE #4) the top cases of largest variety of translations provided by the contributors are invectives ($Dummkopf_n$ (DE #1) and $Trottel_n$ (DE #2)) which are hardly due to their semantic complexity.

Disregarding case, over 5% of the translation pairs are identical, these include proper names but mostly technical loan words, from either of the two languages, or a third, such as evidenced by the EN-DE pairs: $bodyguard_n$ – $Bodyguard_n$, $femoropopliteal_j$ – $femoropopliteal_j$, or $schadenfreude_n$ – $Schadenfreude_n$.

able to assume that relation triples held within translated, aligned phrases ($x_{SL}y_{SL}$ aligned with $x_{TL}y_{TL} \wedge x_{SL}\, R\, y_{SL} \rightarrow x_{TL}\, R\, y_{TL}$)). In our case, we start directly from a weighted triple and take via bilingual lexical translation evidence as sufficient to assume the TL relation holds: $x_{SL}\, R\, y_{SL} \wedge x_{TL} \in tr(x_{SL}) \wedge y_{TL} \in tr(y_{SL}) \rightarrow x_{TL}\, R\, y_{TL}$.

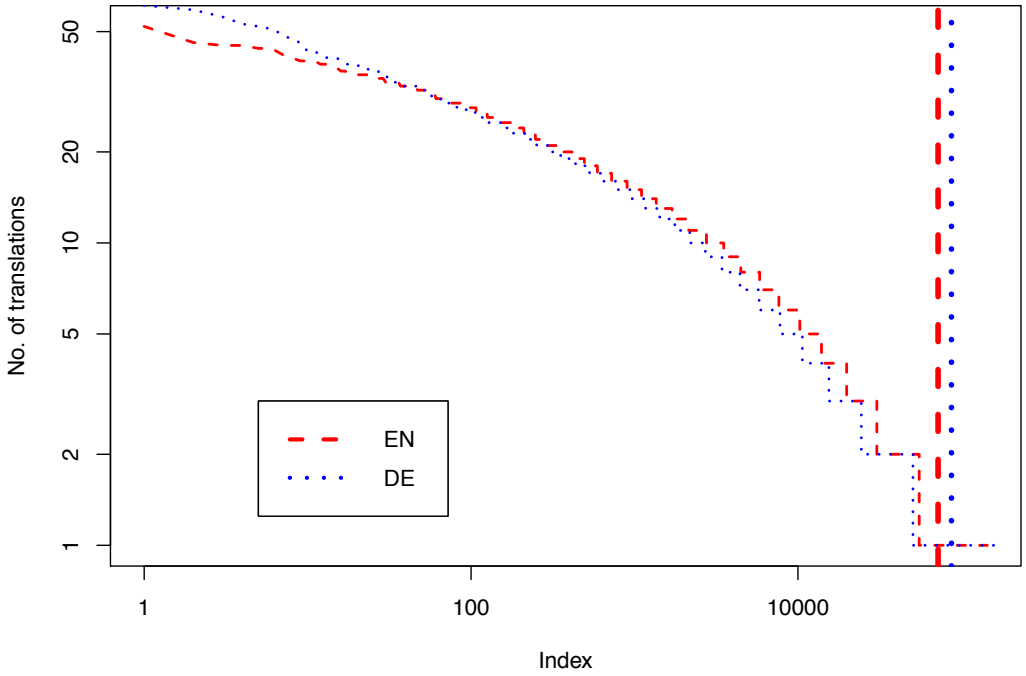


Figure 5.1.: Logarithmic plot of the translational variance in the English-German lexicon dict.cc. The x-axis indicates the index of the words of each language sorted by their number of translations. Vertical lines designate the 50% threshold within the respective vocabularies; for both languages, well over half of the vocabulary has only a single translation.

5.1.2. Formalization of DM as graph

Conceptualizing our SL-DM as a directed graph (such as in Figure 3.2) allows us to construe the task as **graph translation** (Mihalcea and Radev, 2011). Thus a DM can be defined as a triple $G = (V, E, \sigma)$ where V is its set of vertices, the covered subset of SL's vocabulary; E the set of typed edges connecting these words; and $\sigma : V \times E \times V \rightarrow \mathbb{R}^+$, the edge-weighting function. We will use S and T to refer to the source and target language vocabularies, respectively, and index with them to disambiguate as necessary. For example, $G_S = (V_S, E_S, \sigma_S)$ would denote the SL DM. We can describe the process of graph translation using these definitions.

5.1.3. Formalization of DM-graph translation

Building on the assumption that dependency relations are language-independent which – while certainly incorrect in this extreme formulation, does represent an empirically motivated (McDonald, Nivre, Quirmbach-Brundage, Goldberg, Das, Ganchev, Hall, Petrov, Zhang, Täckström, Bedini, Bertomeu Castelló, and Lee, 2013) simplification – we devote our translation efforts completely to the set of vertices, leaving the edge label information untouched.[2]

In the simplest case, a translation lexicon would correspond to a one-to-one mapping between the SL and TL vocabularies $tr : S \rightarrow T, tr(w_S) \mapsto w_T$. Then, the transformation would merely constitute a relabeling. Obviously, this is not the case with natural languages in which many words have the same meaning and most words have multiple meanings. As a consequence, the translation function amounts to a function $tr : S \rightarrow \mathcal{P}(T)$, mapping a word in $w_S \in S$ to a subset of T, i.e. $tr(w_S) = \{w_T^1, w_T^2, \ldots\} \in \mathcal{P}(T)$, an element of the powerset of T.

5.1.4. Performance profile

The goal of pivoting a high-quality SL semantic model into lower-resource TL is to leverage either higher quality SL parses or the larger amount of data to obtain TL models with high performance comparable to a monolingually induced one. There are, however, two distinct aspects of the resulting model in which we can measure this performance: **quality** and **coverage**. These two measures together are what we term the model's **performance profile**.

Quality. This refers to the adequacy of the model which enables – on a particular level of analysis, e.g. the level of individual words or word pairs – us to make correct predictions. This could be measured in accuracy in a classification task or correlation in a relatedness task, for example.

[2] In principle, the approach to translating the lexical nodes could be further extended to include edges which are themselves fully or partially lexicalized, i.e. they contain only lexemes and no grammatical or parser output, while leaving the non-lexicalized or grammatical edges intact. This larger design space for cross-lingual DMs of hybrid node/edge translations lies outside the scope of this thesis.

Coverage. This simply refers to the percentage of the queries to the model on which predictions can actually be made. The original DM was chosen in such a way to cover the items of standard English datasets for lexical semantic tasks (e.g. wordsim, SAT analogy). This was to ensure coverage on those datasets while keeping the size of the model manageable. Coverage is simply a percentage value of covered items which depends naturally on the dataset at hand.

Differences between ML and XL performance profiles. We assume that in general, cross-lingually induced models will show a complementary performance profile to monolingual models (Mohammad et al., 2007; Peirsman and Padó, 2011). Cross-lingual DMs are extracted from higher-resourced SL corpora which almost by definition will be parsed more accurately than TL corpora. This means, monolingual models will suffer from lower quality compared to cross-lingual models. In addition, the translation process can be conceptualized as acting as a further filtering step (cf. Section 5.2), thus optimizing cross-lingual models for higher quality. This comes at the expense of coverage, however, as only those TL words have a chance of being represented in the resulting translated model if they are present in the translation lexicon, which can be viewed as the coverage bottleneck. In contrast, monolingual models – in particular for under-resourced languages – while suffering in quality, will reflect actual SL language use and thus have a higher coverage on most tasks.

5.2. Graph Translation

5.2.1. Ambiguity in Unfiltered Translation

As shown in Figure 5.1, translation on the level of words is not a bijection but rather a many-to-many relation. We model this situation using two functions: $tr : S \rightarrow \mathcal{P}(T)$ which translates source words into sets of target words, and $tr^{-1} : T \rightarrow \mathcal{P}(S)$ which translates target words back into the source language.[3] The simplest way to translate nodes using *tr* is to use all of a node's translations.

[3]Obviously, tr^{-1} is not the inverse of *tr* in a strict mathematical sense; instead, this notation serves simply to model the reverse direction of translation.

Thus, each SL edge between lemmas s_1 and s_2, results in $|tr(s_1)| \cdot |tr(s_2)|$ translated TL edges:

$$E_T = \{(t_1, l, t_2) \mid \exists (s_1, l, s_2) \in E_S \text{ s.t. } t_1 \in tr(s_1) \wedge t_2 \in tr(s_2)\} \tag{5.1}$$

5.2.2. Weighting function

The score σ_T of a target edge can then be defined as some summary statistic of the scores of all SL edges that map onto it. In keeping with the simple approach we adopt the mean of these SL scores as an unbiased and uninformed estimate of the unknown true weight:

$$\sigma_T(t_1, l, t_2) = \frac{\Sigma_{s_1 \in tr^{-1}(t_1), s_2 \in tr^{-1}(t_2)} \sigma_S(s_1, l, s_2)}{|tr^{-1}(t_1)| \cdot |tr^{-1}(t_2)|} \tag{5.2}$$

While applying the **maximum** function might in some cases end up capturing the most credible translation source's score, its sensitivity to outliers would harm the robustness of the model.[4] We take the mean as it is less sensitive to outliers than maximum while still accounting for trends present in the score, as opposed to, e.g. using the median score. In our view, this constitutes a reasonable compromise between integrating scoring information from multiple source edges while at the same time, punishing high degrees of ambiguity or translational variance as less reliable as opposed to having only a few, high-scoring source edges. In addition, unlike e.g. taking the sum total of source scores, it is also normalizes regarding the number of translations, thus penalizing words with many unrelated senses. However, Figure 5.2 indicates the obvious issue with this approach: **over-generation**.

This poses two significant problems. First, the TL graph will contain a much larger number of edges than the SL graph – e.g., using EN-DE dict.cc, the edge $\langle pencil_n \text{ SBJ\_TR } use_v \rangle$ has 72 (the product of each word's translational variance: $|tr(pencil_n)| = 8$, $|tr(use_v)| = 9$) German translations. Second, the quality of the target DM suffers by the introduction of noise in the form of spurious TL edges.

[4] The minimum is clearly also sensitive to outliers; it overeagerly punishes the existence of low-scored source edges which would be unfortunate in many cases.

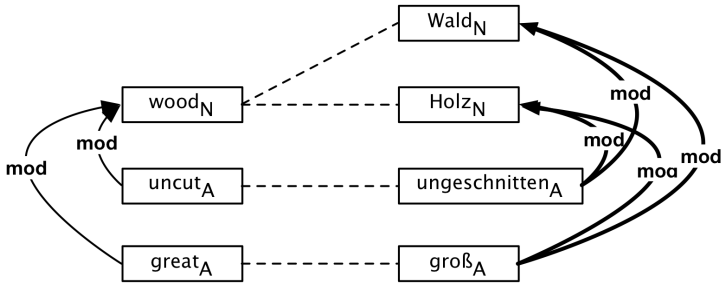


Figure 5.2.: Translating ambiguous nodes in DM graph

In the example of *copse – Gehölz, Hain*, (cf. Figure 5.3) the various translations are synonymous, and Eq. 5.1 is appropriate.

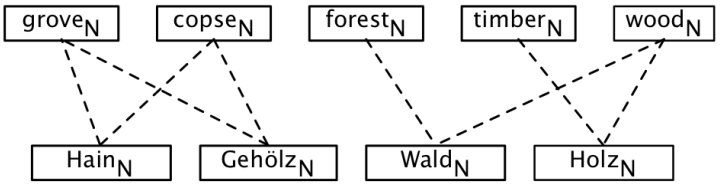


Figure 5.3.: Translation context for $wood_n$

As whichever source edge is used, the semantic (near-)equivalence of the source nodes means the resulting TL edges contain the same semantic information. In many cases, however, the lexical ambiguity of the source terms will lead to TL edges going beyond synonymous relations. Consider as an example, the case of *wood* with its senses as *forest* and *timber* translated as intro German as *Wald* and *Holz*, respectively. The edge translation definition of Equation 5.1 conflates these senses, as Figure 5.2 illustrates.

On the left of the figure we have DM edges between *wood* and two of its adjectival modifiers, *uncut* (being more plausible for the *timber* sense of *wood*) as well as *great* (which is more plausible for the *forest* sense).

The right-hand side shows a subset of the German translations according to Equation 5.1. Both *Holz* (*timber*) and *Wald* (*forest*) are now linked to both adjectives, leading to the presence of unwanted edges in the TL DM.

5.3. Filtering by Backtranslation

In this section, we describe an approach to reducing the effect of ambiguous translation on the TL DM. Important in this undertaking is that we do not introduce any additional requirements of resources – monolingual TL, SL or otherwise – so as to maintain the highest degree of applicability possible.

Since the relationship between translations – i.e. whether particular TL word pairs are to be taken as 'within sense' or not – is not directly evident from the translation lexicon, we will describe a method for exploiting redundancies present in SL DM. The main idea is to leverage quasi-synonymous edges that express the same semantic relation using different words, e.g., $\langle book_n \text{ OBJ } like_v \rangle$ and $\langle novel_n \text{ OBJ } enjoy_v \rangle$, by measuring and scoring target edge candidates by how well they can be **backtranslated** (Somers, 2005) into the source DM. This is similar to the intuition used in the proposed machine translation (MT) evaluation method OrthoBLEU (Rapp, 2009): It improves on the standard ML performance metric BLEU (Papineni, Roukos, Ward, and Zhu, 2002) with its reliance on the availability reference translations. The idea is to use sentence-level backtranslations to assess the quality of a bi-directional MT system by its ability to faithfully reproduce the original input sentence.[5]

5.3.1. Backtranslation applied to DM graph

This idea applied to the situation of translating the DM graph is represented in Figure 5.4. To simplify our example, we will assume that *wood* has two translations, but that *uncut* has only one.

Then the English edge $\langle uncut \text{ MOD } wood \rangle$ translates into two German candidate edges: $\langle ungeschnitten \text{ MOD } Holz \rangle$ and $\langle ungeschnitten \text{ MOD } Wald \rangle$. At this point there is no distinction or preference between the two. However, when backtranslating these candidates, the first, $\langle ungeschnitten \text{ MOD } Wald \rangle$, maps only onto the original edge. We can say there is less evidence for the validity of this TL edge candidate. The second one on the other hand $\langle ungeschnitten \text{ MOD } Holz \rangle$, is

[5] OrthoBLEU does this by calculating the proportion of matching sequences of length 3, thus, the system that is able to exactly reproduce the original input after translating from source into the target language and back, it receives a perfect score of 1.

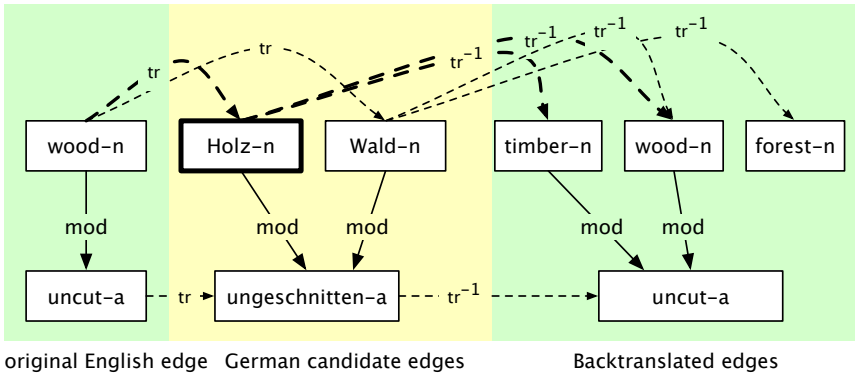


Figure 5.4.: Backtranslation of an ambiguous edge in DM. Translation (tr) and backtranslation (tr^{-1}) are indicated using dashed lines, while DM edges are represented using solid lines. The three colored boxes represent SL/TL/SL boundaries, respectively.

backtranslated into an additionally existing source edge, $\langle uncut_v \text{ MOD } timber_n \rangle$, which can be taken of evidence of the reliability of this TL edge and can serve to discount or dismiss this translated edge from consideration.

The evidence coming from the process of backtranslation is similar allows for a maximum scoring scheme: Trust the highest backtranslation-scored TL edge.

We can then operationalize this insight by adding a condition to Equation 5.1, namely that target edges must be the highest-scoring edges for some source edge.[6]

$$E_T = \{(t_1, l, t_2) \mid \exists (s_1, l, s_2) \in E_S \text{ s.t. } t_1 \in tr(s_1) \wedge t_2 \in tr(s_2) \\ \wedge \sigma_T(t_1, l, t_2) = max_{t \in tr(s_1), t' \in tr(s_2)} \sigma_T(t, l, t')\}$$

where $\sigma_T(t, l, t')$ is the TL edge score as defined in Equation 5.2.

It should be noted that this filtering scheme can be made to accept fewer or more edges by loosening or tightening this constraint allowing in larger or smaller numbers of target edges that a source edge can translate to. A more liberal variant

[6]Recall that our target scores σ_T have already been defined in terms of source edge scores, so no redefinition of the scoring function is necessary for the TL DM.

could, for instance, take the top n-scored TL edges, whereas a stricter variant could, altogether abstain from translating a source edge if no unique best edge exists. This gives a lot of freedom to those interested in constructing their own DM, allowing for the fine-tuning of its characteristics, e.g. whether higher coverage or accuracy is to be given preference.

Translation graph components. Taking the bilingual dictionary as a bipartite graph, the question arises what the connectivity of that graph is and how large its connected components are.

Figure 5.5 graphs the connected component size, i.e. the number of words n (EN and DE) in a component, against the number of components of that size freq(size $= n$). The blue line plots the cumulative coverage corresponding to each component size, i.e. at each component size n, the vertical position is at $\frac{\sum_{n' \leq n} n' \cdot freq(\text{size}=n')}{|\text{vocab}_{EN} \cup \text{vocab}_{DE}|}$ as a percentage value.

The first insight is that approximately half of the vocabulary is located in small components of size ≤ 4. The next 10% of the vocabulary is in moderately sized components ($5 \leq$ size ≤ 124), with the final 40% falling into unconnected large components of sizes: 14K, 30K, 90K.

One might consider different restriction schemes on the permissible component sizes as an additional level of constraint to such a backtranslation filtering approach to cross-lingually inducing DMs.

5.4. Qualitative Analysis of the filtered $\text{DM}_{\text{EN}\rightarrow\text{DE}}$

Graph density. Using the English DM made available by Baroni and Lenci (2010)[7], the filtered $\text{DM}_{\text{EN}\rightarrow\text{DE}}$ has a vocabulary size of 60K and 25K link types and a density of $8.78 \cdot 10^{-6}$. In order to compare it more directly with the above mentioned DMs we restrict it to its top 30K words[8] which increases the density to $3.04 \cdot 10^{-5}$ – around twice the density of the restricted DM_{DE}. This is most likely due to the translational variance introducing noisy edges.

[7] Available at: `http://clic.cimec.unitn.it/dm/`.

[8] If we look at SDEWAC's top 30K words, only around 60% of these are covered by $\text{DM}_{\text{EN}\rightarrow\text{DE}}$; an effect of the lacking coverage of the bilingual dictionary.

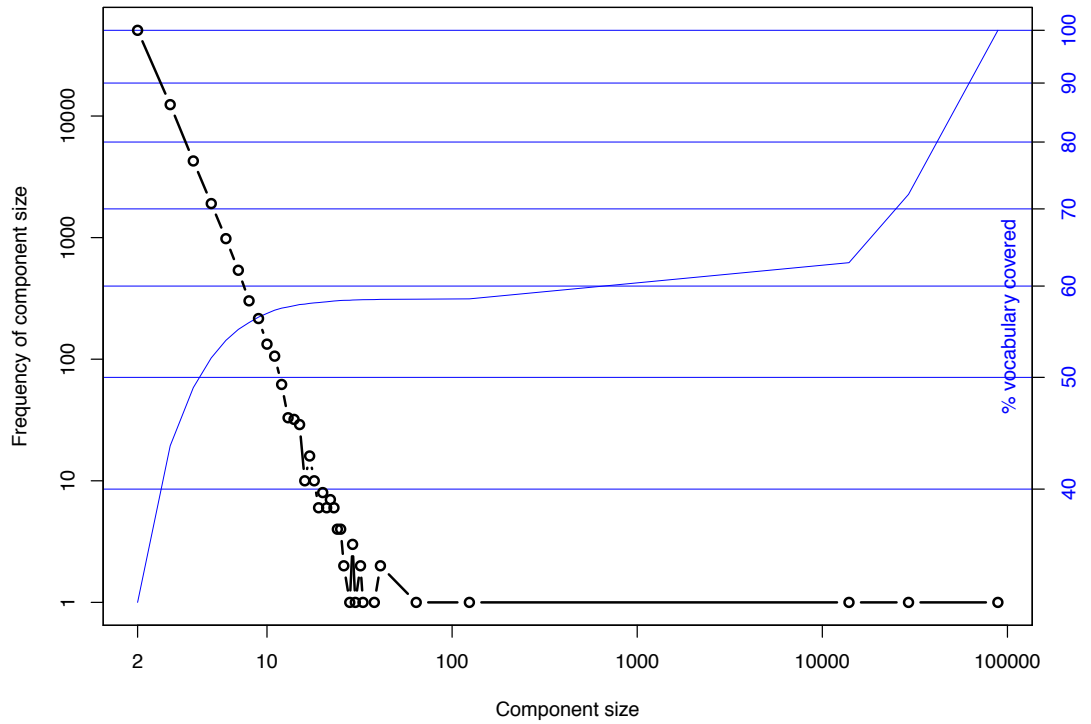


Figure 5.5.: Connected components in `dict.cc` graph. Black: component size vs. component size frequency. Blue: cumulative coverage of EN & DE words.

Analysis of W×LW vectors. Comparing these values Table 4.4 for the top 30K words, we notice a number of differences. Most strikingly, the number of edges for the cross-lingual model is drastically higher than for DM_{DE}. On average, a word in the $DM_{EN \to DE}$ graph has 40 times the number of edges as in DM_{DE}, with a noticeable trend across the frequency bands (high: 20x, mid: 25x, low: 70x).[9]

With regards to vector norms the differences decrease, only an average of 2.5 higher norms for word vectors in the cross- versus the monolingual model. At the same time, the norms for the highest frequency bands are almost at parity.

[9]One possible explanation for the dramatically higher number of edges for low-frequency words is that there is only a small correlation between frequency and translational variance – the main factor in the number of edges: $r = .08$ ($p < .001$).

	high	mid	low	across
# Edges per word				
min	20711	5814	216	
median	37774.500	11277.500	2013	
mean	53556.880	12167.530	2359.440	
max	234972	20341	5718	
Word vector L^2 norm				
min	118.501	80.876	62.745	
median	3380.174	1362.404	603.409	
mean	4003.256	1563.619	635.194	
max	15131.073	7147.235	1611.644	
Cosine similarity				
min	0.000	0.000	0.000	0.000
median	0.042	0.002	0.019	0.004
mean	0.056	0.014	0.028	0.020
max	0.724	0.635	0.585	0.724
non-zero sims	72.1%	54.6%	92.4%	72.5%

Table 5.2.: $DM_{EN \rightarrow DE}$ W×LW stats on top 30K words.

The similarities show a similar trend, the highest factor for low-frequency words (8x) and the lowest for more frequent words (3x) while the mid frequency range is more or less at parity. Here we suspect the effects of larger norms and increased sparsity (via larger number of edges, i.e. components) balance one another out.

We have shown how with very straightforward and simple means a dense TL DM can be constructed. In the following chapters we will see that these cross-lingual SDSMs perform on par with much more involved models.

Part III.

Evaluating and Combining SDSMs

Chapter 6.

Evaluation of single source-language DMs

In this chapter, we evaluate and compare the performance of our monolingually and cross-lingually induced distributional memories on the battery of tasks presented in Chapter 3. As there are a number of design choices that can be made for our models, we argue in the following section for one particular formulation of our DMs. In presenting the results, we follow the diagram of increasing semantic and syntactic complexity involved in the task, as sketched in Figure 3.5, by moving from evaluations of lexical relatedness to multiple argument composition.[1] While the tasks discussed can be viewed as in-vitro tests, they are widely used for model selection in work on distributional semantics and as a result, we can compare the results of our models to prior work. Additionally, our focus is not to project the effect of any model type on the multitude of possible downstream tasks. To the contrary, our aim is to make the case for expanding the reach of the already proven general-purpose SDSMs, Distributional Memories, in terms of the number of languages whose resources can be supplemented with such advantageous models by showing how our induction methods yield such general-purpose models for novel target languages.

[1] Note that our hope is not necessarily for our models to beat custom-built models but rather to show the flexibility and broad applicability of our DMs. If our models can perform comparably well to those designed specifically for the tasks at hand, we consider this a success.

6.1. Task 1 – Word relatedness

Reminder of task. The goal of this lexical relatedness task is to test how well a model can predict the degree of relatedness of German word pairs. The Gur350 dataset described in detail in Chapter 3 includes closely related, somewhat related, and unrelated word pairs ranked on a five-point Likert scale between 0 (unrelated) and 4 (synonymous) by German native speakers. It contain nouns, verbs and adjectives.

Reminder of evaluation. The quality of the predictions is obtained on the whole dataset as a measure of correlation (Pearson's r) between predictions and judgments. In cases where a word in a pair is not covered by the model we assign a default prediction of 0.[2] This corresponds to saying that an uncovered word pair should be expected to have a relatedness as picked randomly from the covered items in the dataset. We measure our models' prediction in terms of coverage and correlation on covered.

DM models. First we start with simple first-order co-occurrence models. These are models that do not make use of the distributed nature of the representations, instead using only association scores coming from the weighted edges between the nodes, i.e. words, in the DM graph.[3] This means they only give estimates to word pairs which have been directly observed in the underlying corpus, i.e., the words have appeared in the context of one another. They thus capture syntagmatic relationships between words as opposed to the paradigmatic relationship captured by word similarity in a structured word space. The intuition behind these models is that if there is a high level of association present in a single edge between words, then these values themselves might match well with human predictions of semantic relatedness. We introduce two first-order models:

[2] We also ran experiments using the average relatedness predicted by model to all covered word pairs in the dataset. No substantial difference in performance was found in either DM induction type.

[3] They are, in principle, similar to the association metrics used by Michelbacher et al. (2007) for collocation recognition in that they are asymmetric in nature. However, the scores in our models are not computed as probabilities and take many relation types into account.

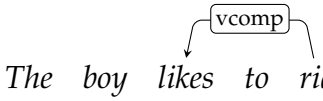
$\mathcal{M}_{\text{highest}}$: assign the highest σ score in the DM:

$$\pi_{\mathcal{M}_{\text{highest}}}(w_1, w_2) = arg\,max_{\text{LINK}}\sigma(\langle w_1 \text{ LINK } w_2\rangle)$$

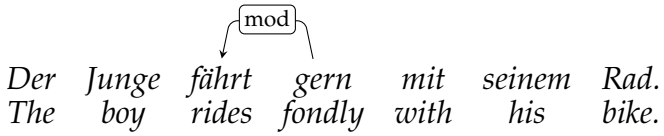
taken over all available links between the two given words in the DM.

$\mathcal{M}_{\text{gram}}$: read off DM σ scores for $\pi_{\mathcal{M}_{\text{gram}}}(w_1, w_2) = arg\,max_{link}\sigma(\langle w_1 \text{ LINK } w_2\rangle)$ for any grammatical, i.e. non-lexicalized link: subjects in an intransitive or transitive construction (SUBJ_INTR, SUBJ_TR); direct or indirect objects (OBJ, IOBJ); subject-object pairs linked by some verb (VERB), verbal complements (VCOMP); or noun modifiers (NMOD).

The motivation for considering the highest associated link between two words is that higher associations can be deemed more relevant: either in the case of gathering the judgments, as incorrect or less likely analyses – e.g. transitive as opposed to intransitive constructions (van Gompel, Pickering, Pearson, and Jacob, 2006) – are still active in the mind; which could provide a predictive edge over $\mathcal{M}_{\text{gram}}$ in the monolingually induced DM. In addition, there is reason to believe that in the case of the cross-lingually induced DM the most salient relation between a head and a dependent might be coded by a different link. Consider the following English sentence as example:



In German this scenario would most likely be rendered as:



Given our simple DM translation scheme which keeps the links constant we would in this case, due to the shift in link types from English to German, expect to find plausible associations in a quite distinct link type. This is similar to the approach taken by Peirsman and Padó (2011), who defined their nearest-neighbor based cross-lingual plausibility model using an *arg max*over relations to select the best translation of the supplied relation given the predicate and argument.

Model $\mathcal{M}_{\text{wordsim}}$	r_{cov}	cov
DM_{DE} (AllL)	.43	.60
DM_{DE} (SPrfL)	.43	.60
$\text{DM}^{\text{filter}}_{\text{EN}\rightarrow\text{DE}}$ (AllL)	.42	**.61**
$\text{DM}^{\text{filter}}_{\text{EN}\rightarrow\text{DE}}$ (SPrfL)	**.49**	.49

Table 6.1.: Task 1: Results of including only selectional preference-based links into vector construction and thus relatedness calculation. This corresponds to a restriction of model to a subgraph (or filtering out 'link-slices' in the tensor) resulting in a compressed model. (Highest values in each column are in boldface.)

We also test our monolingual DM (DM_{DE}) as well as our cross-lingually induced DM ($\text{DM}_{\text{EN}\rightarrow\text{DE}}$) using the word vector-based W×LW models:

$\mathcal{M}_{\text{wordsim}}$: For each word pair $\langle w_1\ w_2\rangle$, we calculate simple vector-based word similarity in the W×LW space; the prediction of the model is thus:

$$\pi_{\mathcal{M}_{\text{wordsim}}}(w_1, w_2) = cos\, sim(\overrightarrow{w_1}, \overrightarrow{w_2})$$

where each word vector is obtained via Equation 3.2.

Recall that nodes in DM contain both incoming as well as outgoing edges. Our monolingually induced DMs by default use all of the available information in the relations. However, it is not clear that in a cross-lingual setting all information is equally reliable. The intuition in this case is that selectional preferences would be the most informative type of relational information and should be most likely to survive the translation process. With regard to verbs, for example, we can expect knowledge about their arguments to be more informative than the relationships with their governors. By a similar token, nouns would be more robustly characterized by the verbs that link to them than, say, their modifiers. We tested this idea by computing semantic relatedness between word vectors either on full vectors (covering all link types; condition 'AllL') or on a filtered version that includes only inverse links for verbs and regular links for nouns and adjectives (condition 'SPrfL'). The results bore out our hypothesis: In the monolingual setting, we find

there is almost no difference between AllL and SPrfL. Thus, we adopt the AllL modality for DM_{DE}. In contrast, we note a distinct quality-coverage trade-off in the cross-lingual results, with a higher accuracy for SPrfL – even higher than DM_{DE}. For this reason, we will adopt (SPrfL) for all cross-lingual DM experiments.

It is worth noting that Table 6.1 shows that with our cross-lingual models $DM_{EN \to DE}$ we are able to achieve the performance of a specialized target language model, DM_{DE}, while not requiring any target-language corpus data. All we required was the English DM and a translation lexicon of around 300K translation pairs. Then, by applying the filtering, we even improve the quality of the predictions, albeit at the cost of losing coverage of some items. In Chapter 7, we will show how it is possible to combine these two model types to take advantage of higher accuracy for one model and higher coverage for the other.

Non-DM models. We also consider non-DM models that exemplify either a lower level of resource requirements, such as DSMs, or best-performing custom models from the literature.

> BOW a bag of words model over SDEWAC with a context size of 10:
> $\pi_{BOW}(w_1, w_2) = \cos(\vec{w_1}, \vec{w_2})$

This is a standard DSM implementation that represents the baseline for what an uninformed distributional model will capture.

> $BOW^{\text{dim-red.}}$ is a dimensionality reduced version of the BOW model, it has had principle component analysis (PCA) performed to reduce the numbers of dimensions to 500.

As mentioned in Chapter 2, using dimensionality reduction methods can help reduce the noise in a DSM. Conceptually, these methods can discover and factor out latent semantic patterns (cf., e.g., Landauer and Dumais (1997a)) making the basis vectors of the resulting space maximally distinct from one another, thus warranting their orthogonality which in terms of semantic relatedness measures such as cosine similarity renders them totally unrelated.

The next models were first described by Budanitsky and Hirst (2001) building on work by Resnik (1995), J. Jiang and W. Conrath (1997) and Lin et al. (1998) on

information-theoretic models of semantic similarity. Their performance on the Gur350 dataset was reported by Zesch et al. (2007).[4]

Lin$_{GN}$: uses the Lin-similarity measure in the taxonomy of GERMANET to determine relatedness.

$$\pi_{\text{Lin}_{GN}}(w_1, w_2) = 2 \cdot \frac{IC(w_1 \vee w_2)}{IC(w_1) + IC(w_2)}$$

where IC is the **information content** of a synset[5] and $w_1 \vee w_2$ denotes the least common subsumer – i.e. the most specific common ancestor – of words w_1 and w_2 in the taxonomy hierarchy.

JC$_{GN}$: (J. Jiang and W. Conrath, 1997) uses conditional probability starting from the assumption of being at the least common subsumer:

$$\begin{aligned}\pi_{\text{JC}_{GN}}(w_1, w_2) &= \log\left[p(w_1|w_1 \vee w_2)\right] + \log\left[p(w_2|w_1 \vee w_2)\right] \\ &= IC(w_1) + IC(w_2) - 2 \cdot IC(w_1 \vee w_2)\end{aligned}$$

This follows from the definition of conditional probability:

$$\log\left[p(a|a \vee b)\right] = \log\left[\frac{p(a \wedge (a \vee b))}{p(b)}\right] = \log\left[p(a \wedge (a \vee b))\right] - \log\left[p(b)\right]$$

and since a is always under any of its subsumers, i.e. $\forall x.a \leq a \vee x$, it follows $p(a \wedge (a \vee b)) = p(a)$ and thus:

$$\log\left[p(a|a \vee b)\right] = \log\left[p(a)\right] - \log\left[p(a \vee b)\right].$$

The intuition behind this model is that when w_1 and w_2 are good candidates for the

[4]Other models using monolingual and cross-lingual approaches to predicting semantic relatedness on the Gur350 dataset were described by Mohammad et al. (2007), but as their performance was provided without coverage values we omit them from our considerations here.

[5]The information content measures how unlikely it is to sample an instance of a given concept (i.e. a descendent): $IC(w) = -\log(p(w))$. E.g. since the root of the taxonomy subsumes all other concepts, its probability is $p(\text{root}) = 1$ leading to the minimum information content value of $IC(\text{root}) = 0$. A concept that subsumes no others of a total of n concepts would have $IC(c) = -\log(p(c)) = -\log(\frac{1}{n}) = \log(n)$, becoming more informative the more concepts there are.

concept $w_1 \vee w_2$ then their large conditional probabilities $p(w_i|w_1 \vee w_2)$ $(i \in \{1, 2\})$ lead to higher relatedness via the monotonic logarithmic transformation. Also, the lower these probabilities become the stronger they are penalized, with probabilities close to 0 leading to similarities close to $-\infty$.

$\text{JC}_{\text{GN}} + \text{PL}_{\text{WP}}$: is a simple linear combination of the above model with the other best performing model in Zesch et al. (2007):

$$\pi_{\text{JC}_{\text{GN}}+\text{PL}_{\text{WP}}}(w_1, w_2) = \pi_{\text{JC}_{\text{GN}}}(w_1, w_2) + \pi_{\text{PL}_{\text{WP}}}(w_1, w_2)$$

where the second model PL_{WP} (Gurevych, 2005) takes the minimum path length between the words' disambiguated articles on Wikipedia in its hierarchy of concept categories.[6]

Baseline.

Freq: A frequency baseline that predicts higher relatedness for pairs whose constituent words are more frequent.

In psycholinguistics, it is well known that the frequency of a word has an effect on the outcome and speed at which judgments are given in experimental tasks (Broadbent, 1967; Taft, 1979). One possible reason for this observation is that higher frequency words are connected in some manner to many more words leading to an increase in the rate of activation – coming from a more densely connected network of association.

$$\pi_{\mathcal{M}_{\text{Freq}}}(w_1, w_2) = min(\text{freq}(w_1), \text{freq}(w_2))$$

The predictions of the model are the minimum corpus frequency in SDEWAC between the two words. The intuition behind this is that being on the lower end of the frequency range will have a more significant impact on the relatedness by

[6]These are the pages listed in the footer information of a Wikipedia article that structure the single article pages into larger groups, e.g. the article for *correlation* is listed under the two categories *Covariance and Correlation* and *Dimensionless Numbers*. They serve many purposes organizing social groups, events throughout history but also taxonomic information in the domains of biology and even beyond, but generally quite sparsely populated.

being further outside this assumed dense center of higher activation for higher frequencies.

Results. We first devote our attention to the many DMs and model types in Table 6.2. Here we see that the first-order models perform worse than using the full vectors to calculate relatedness. This means the distributional information provides more help in discovering the relatedness of words over trying to find direct links between the words. This also explains the lower coverage for these first-order models: while the $\mathcal{M}_{\text{wordsim}}$ models rely on second-order co-occurrences – i.e. shared contexts in the corpus, which could evidence topical or grammatical overlap – the $\mathcal{M}_{\text{highest}}$ and $\mathcal{M}_{\text{gram}}$ models require the words to have been seen in the same context and connected via a link pattern covered by the DM. Since the DMs as W×LW spaces capture the information well it appears that the semantic relations in the dataset are more complex than could be quantified by a single statistical association value in an edge weight.

DM type	r_{all}	r_{cov}	$cov.$
First-order models – $\mathcal{M}_{\text{highest}}$			
DM_{DE}	.10	.13	.40
$\text{DM}_{\text{EN}\rightarrow\text{DE}}$	.14	.17	.34
First-order models – $\mathcal{M}_{\text{gram}}$			
DM_{DE}	.09	.13	.31
$\text{DM}_{\text{EN}\rightarrow\text{DE}}$	.10	.11	.19
W×LW models – $\mathcal{M}_{\text{wordsim}}$			
DM_{DE}	**.38**	.43	**.60**
$\text{DM}^{\text{naive}}_{\text{EN}\rightarrow\text{DE}}$	.28	.45	.49
$\text{DM}^{\text{filter}}_{\text{EN}\rightarrow\text{DE}}$	.33	**.49**	.49

Table 6.2.: Comparison of structured models on Gur350 (highest values in each column are in boldface).

We also find that $\text{DM}^{\text{naive}}_{\text{EN}\rightarrow\text{DE}}$ performs worse than $\text{DM}^{\text{filter}}_{\text{EN}\rightarrow\text{DE}}$ and as a result, in all following tasks and analyses we will only consider $\text{DM}^{\text{filter}}_{\text{EN}\rightarrow\text{DE}}$ (for which we use the simplified notation $\text{DM}_{\text{EN}\rightarrow\text{DE}}$). Additionally, it also has a significantly smaller size allowing for the faster calculation of predictions.

We also find a pattern of higher coverage for the monolingual DM_{DE} models and – except for $\mathcal{M}_{gram}$ – higher correlation values for the cross-lingual models. One issue affecting the performance of $DM_{EN \to DE}$ in the $\mathcal{M}_{gram}$ modality is that the more selectively we look at the presence, absence or weighting of edges between two words in the target language DM, the less reliable the information located there becomes. Taken in aggregate, e.g. with $\mathcal{M}_{wordsim}$, or even a global view over all links such as in $\mathcal{M}_{highest}$ helps in the cross-lingual DM, as compared to the monolingual one. The First-order models are overly selective: They rely upon a single data point to model the variance of semantic relatedness present in the dataset.[7]

Model	r_{all}	r_{cov}	$cov.$
Baseline & non-DM models			
Frequency	.14	.14	**.97**
BOW	.20	.21	**.97**
$BOW^{dim\text{-}red.}$	.34	.37	**.97**
Lin_{GN}	NA	.50	.26
JC_{GN}	NA	.52	.33
$JC_{GN} + PL_{WP}$	NA	**.59**	.33
W×LW wordsim models			
DM_{DE}	**.38**	.43	.60
$DM_{EN \to DE}$	.33	.49	.49

Table 6.3.: Task 1: Coverage and correlation (Pearson's r) for predicting word relatedness on the Gur350 dataset. (Highest values in each column are in boldface.)

Next we compare our best models against the baseline and literate models in Table 6.3. The frequency model, along with the unstructured DSMs (cf. first three rows in Table 6.3), have almost perfect coverage – missing only such uncommon words as *Inaugurationsmesse* (inaugural exhibition), *Geschirrdurcheinander* (disarray of dishes) and *Volierenzelt* (aviary tent). However, the correlations achieved by these models are relatively low: .21 for BOW and .37 for $BOW^{dim\text{-}red.}$. For the frequency model this is due to the fact that no direct semantic information linking the two words is captured; instead, only effects of being more or less common

[7] Evaluating the predicted word similarities using Spearman's ρ as opposed to Pearson's r did overall improve the performance profile of our models but in order to compare against results from the literature, we report standard Pearson correlation values.

in the corpus is exploited as a potential correlate of relation. At the same time, dimensionality reduction obviously helps improve the quality of the BOW space immensely.

Both SDSMs DM_{DE} and $DM_{EN \to DE}$ have lower correlation values (.43 and .49) than the best models from the literature (.59 for JCGN and .50 for Lin_{GN}). latter models make use of labor-intensive semantic resources such as GermaNet which makes them less available to novel languages. As a result of having this additional resource requirement they also suffer in coverage making predictions for only one quarter to one third of the total number of word pairs in the dataset. In general, lexical taxonomy-based models (e.g. WORDNET, GERMANET) will only be able to make predictions of semantic relatedness or similarity between words in the same hierarchy, i.e. for which there exists a path between them, which, among other things, significantly reduces their ability to relate words across parts of speech.

We compare the best performing models using a one-tailed Fisher's z-test with Bonferroni correction. Accounting for the differences in coverage, we find that the simple DSMs are hard to beat. The only models that are significantly outperformed by any of the top models ($p < .05$) are the frequency baseline and BOW. Also, only $JC_{GN} + PL_{WP}$ outperforms BOW with all other models besides BOW outperforming the frequency baseline. If we choose a significance level of $p < .001$, then the only models to outperform the frequency baseline are the two JC_{GN} models and $DM_{EN \to DE}$. Thus we can say that our structured models perform on par with models using significantly higher resource requirements and whose design (in the case of processing Wikipedia to resolve lexical ambiguities) introduces a number of heuristic choices, with $DM_{EN \to DE}$ doing particularly well. We also note that the simple distributional context information in the DSMs is sufficient for simple lexical relatedness.

6.2. Task 2 – Synonym choice

Reminder of task. The dataset for the task of identifying synonymy has the format:

$$\langle \textit{base}, \textit{candidate}_1, \ldots, \textit{candidate}_4 \rangle$$

where one of the four candidates is synonymous to the target word.

Whereas the goal in the previous lexical relatedness task was to predict an essentially continuous semantic relatedness value, here the task can be described as a relation-based categorical classification. The target is assumed to have a specific relation – synonymy – to one item in a list of possible candidates.

Reminder of evaluation. We evaluate analogously to Mohammad et al. (2007), defining a type of weighted accuracy via scoring (cf. Equation 3.4). In the absence of ties, we obtain a standard accuracy value, i.e. the proportion of correctly determined synonym targets among the items. Otherwise, we obtain a weighted sum of correct predictions, with discounts for ties: $acc_s(\mathcal{M}, \mathcal{D}) = |\mathcal{D}|^{-1}(A + \frac{1}{2} \cdot B + \frac{1}{3} \cdot C + \frac{1}{4} \cdot D)$.

Note that acc_s does not take the actual number of covered items into account. To remedy this, Mohammad et al. also define a precision-inspired acc_c measure defined as the average acc_s per covered item: $acc_c(\mathcal{M}, \mathcal{D}) = \frac{acc_s(\mathcal{M},\mathcal{D})}{cov(\mathcal{M},\mathcal{D})}$, where $cov(\mathcal{M}, \mathcal{D}) \in [0, 1]$ is the coverage of the dataset $\mathcal{D}$ by the model $\mathcal{M}$, i.e. the proportion of items for which a predication can be made.[8]

We also can perform significance tests in this classification task using χ^2 for all models since we do not require the actual predictions from the models to determine significance.

DM models. As in Task 1, we use the first-order models $\mathcal{M}_{\text{gram}}$ and $\mathcal{M}_{\text{highest}}$ which simply use the DM edge weights between the two words' nodes as indicators of association, which now is interpreted as a measure of how close to synonymous the words are. The idea being that words with closely related meanings might reasonably be expected to co-occur in text. We also use $\mathcal{M}_{\text{wordsim}}$ to obtain word similarities in W×LW. For each target, we compute the cosine of its vector with each candidate's vector and assign this similarity to the candidate. Our single-word model ($\mathcal{M}^{\text{word}}_{\text{wordsim}}$) excludes those problems that include phrases as candidates. DM does not provide a means of representing phrases,

[8] It follows from the definitions of $cov = \frac{\#covered}{|\mathcal{D}|}$ and of $acc_s = \frac{score}{|\mathcal{D}|}$ that since the score is always less than the size of the dataset, that: $acc_c = \frac{acc_s}{cov} = \frac{score}{\#covered}$ and $0 \leq acc_c \leq 1$.

instead its association values are indexed by words and the relations described in Chapters 2 & 4.

We also implement a simple model for phrases ($\mathcal{M}^{\text{phrase}}_{\text{wordsim}}$) for the large number of phrasal candidates in the dataset ($538/984 \approx 55\%$): It defines the similarity between a target word and a phrasal candidate as the maximum similarity of the target word and any of the constituent words of the phrase. This can be viewed as assuming that the distinguishing meaning of a phrase is determined solely by that constituent word that is most similar to the target word. Finally, we combine the two models into a single model, $\mathcal{M}_{\text{wordsim}}$. This can be achieved trivially, since the two models $\mathcal{M}^{\text{word}}_{\text{wordsim}}$ and $\mathcal{M}^{\text{phrase}}_{\text{wordsim}}$ partition the dataset $\mathcal{D}$.

It is important to note that the definition of coverage in this task depends on multiple words for each item, namely the target word and the candidate words or phrases.We experimented with two modalities for coverage using DM_{DE}: The first modality (α) excludes items for which the similarity of at least one candidate phrase or word cannot be estimated. This model is conservative in that it abstains from making predictions for items where sparsity or coverage problems have occurred. The second model modality β excludes only those items in which all of the candidates are out of coverage or have zero similarities to the target. Tests showed a significantly higher coverage for β over α (.84 vs. .41) on the RDWP dataset with only minimal effect on performance measured in accuracy on covered (.53 vs. .57). We thus adopt the β modality for our evaluation.

Non-DM models. Mohammad et al. (2007) evaluate a number of monolingually induced models based on GermaNet (Henrich and Hinrichs, 2010; Hamp and Feldweg, 1997), a large lexical resource for German similar to WordNet, falling into two categories. First, gloss-based models which are variants of the Lesk algorithm (Lesk, 1986) applied to the glosses attached to nodes in GermaNet (Gurevych, 2005):

> HPG: *Hypernym pseudo-gloss*, models similarity using n-gram overlap with unambiguous word senses of hypernyms (path length ≤ 3) as glosses
>
> RPG: *Radial pseudo-gloss*, same as HPG except: all semantic relations but hyponymy with a path length ≤ 3 with a distance ≥ 2 to the most abstract root concept.

Second, hierarchical models which derive similarity scores using graph-based similarity measures along the paths in the hierarchical structure of GermaNet.

Baselines. A frequency baseline assigns to each candidate a score equal to the highest frequency of its words. The candidate with the highest word frequency is predicted to be the correct one:

$$\pi_{\mathcal{M}_{\text{Freq}}}(\langle base, candidate_1, ...\rangle) = arg\,max_{i\in\{1,2,3,4\}} freq(candidate_i)$$

Standard bag-of-words (BOW) DSMs are also listed as baseline models. As in the case of our W×LW SDSM models, their predictions are based on word similarity but in a space whose vectors are derived from unstructured co-occurrences. The dimensionality-reduced BOW$^{\text{dim-red.}}$ additionally had a standard matrix factorization technique (principal component analysis, PCA) applied to decrease the number of dimensions to 500. Finally, a random baseline is included which, having a 1 in 4 chance of predicting the correct candidate, trivially corresponds to a 25% accuracy.

Results. In order to assess significance between model performances, we apply a Bonferroni-corrected χ^2 test which takes both the accuracy and coverage into account. We again first look at how the DM models perform overall, then compare among different model types.

Table 6.4 shows the results on the RDWP dataset for choosing synonyms. The first noticeable difference between the first-order models and W×LW models is in coverage. The first-order models overall have much lower coverage scores compared to the $\mathcal{M}_{\text{wordsim}}$ models.[9] This is to be expected as we are requiring the word pairs we compare for scoring to be directly connected, which is quite restrictive. In the cases where we do have coverage, and in contrast to the results in Task 1, we find that the predictions are quite reasonable: acc_{cov} values are between .63 and .70 for first-order models, whereas the covered accuracy of the $\mathcal{M}_{\text{wordsim}}$ models is between .53 and .61. While this is higher than expected, a look at the most common lexicalized link types linking the synonyms reveals a pattern:

[9] We can expect lower coverage overall for this dataset due to its design which was to test the readers' knowledge of uncommon words.

Link type	Word pairs involved in correct synonym detection (DM_{DE})
BEDEUTEN (to mean)	*Authentizität–Echtheit* (authenticity–realness), *Meteorologie–Wetterkunde* (meteorology–weather study), *Priorität–Vorrang* (priority–precedence), *Ironie–Spott* (irony–ridicule)
DARSTELLEN (to represent)	*Honorar–Vergütung* (honorarium–emolument)
VERURSACHEN / ERZEUGEN (to engender)	*Schwankung–Fluktuation* (deviation–fluctuation), *Teilchen–Partikel* (particle)
GARANTIEREN / BIETEN (to guarantee/offer)	*Prosperität–Wohlstand* (prosperity–wealth), *Biotop–Lebensraum* (biotope–living space)
SCHÜTZEN (to protect)	*Tonsur–Kopf* (tonsure–head)
ALS (as)	*Frau–Lesbe* (woman–lesbian)

Here we can see that we can actually find links that directly encode the type of information required to identify synonyms (*mean, represent*). Additionally, links can express underlying causal relationships (*engender, guarantee, offer*) or possible closely related words (protect, as). In the case of $DM_{EN \to DE}$, the most common link types involved in edges that allow for the correct synonym to be detected are from among the overall most frequent grammatical (NMOD and COORD) and lexical link types (OF and BE).[10] The most common grammatical links for both models which lead to successful synonym detection are NMOD, VERB and COORD.

One interesting effect that is distinct from the last task is that the coverage is significantly higher by a factor of approximately 2 for $DM_{EN \to DE}$ than DM_{DE} among the first-order models. We found that this is due to the dataset design: The target words were chosen to be a challenge for Reader's Digest readers, and thus there is a preponderance of loan words. These also happen to often be more or less identical across English and German, e.g. *Bagatelle - bagatelle*, *Generosität - generosity*, *Couscous - couscous* and *Präferenz - preference*. In the W×LW vector spaces, since we do not require direct links in the DM, we have increased ability to make predictions and thus a higher coverage. Also, requiring grammatical links necessarily leads to

[10] Although these lexical links do not clearly show the potentially synonymous relationship between the words – as was the case of those in DM_{DE} – we take the high acc_{cov} values for $DM_{EN \to DE}$'s first-order models as an indication that the edge information in our XL DM is high quality.

Model	acc	acc_{cov}	cov
First-order models - $\mathcal{M}_{\text{highest}}$			
DM_{DE}	.11	.65	.17
$\text{DM}_{\text{EN}\rightarrow\text{DE}}$	.19	.63	.31
First-order models - $\mathcal{M}_{\text{gram}}$			
DM_{DE}	.06	**.70**	.08
$\text{DM}_{\text{EN}\rightarrow\text{DE}}$	.11	.68	.16
W×LW models - $\mathcal{M}_{\text{wordsim}}$			
DM_{DE}	**.48**	.53	**.84**
$\text{DM}_{\text{EN}\rightarrow\text{DE}}$	.46	.61	.58

Table 6.4.: Performance of SDSMs on Task 2: synonym choice. (Highest values in each column are in boldface.)

lower coverage which explains the decrease in coverage between corresponding $\mathcal{M}_{\text{highest}}$ and $\mathcal{M}_{\text{gram}}$ models.

Testing differences in accuracy values between these DM models and the random baseline using a χ^2 significance test, we find that the first-order models perform worse than the $\mathcal{M}_{\text{wordsim}}$ models as well as the random baseline of .25 ($p < .001$). Their coverage is simply too low. Of the W×LW models, DM_{DE} outperforms $\text{DM}_{\text{EN}\rightarrow\text{DE}}$ ($p < .001$) and random.

Among the non-DM models, the frequency and $\text{BOW}^{\text{dim-red.}}$ models dominate the models from the literature (cf. Table 6.5). The difference in coverage (.45 versus .22) and accuracy (.52 versus .44) puts Lin_{dist} ahead of HPG ($p < .05$) and JC ($p < .001$), respectively. No significant difference was found between Lin_{dist} and RPG. Both baselines, Frequency and Random, outperform JC, HPG ($p < .001$) and RPG ($p < .05$). Random does better than JC, HPG ($p < .001$) and RPG ($p < .05$) while Frequency beats them all at a higher level of significance ($p < .001$) as well as Lin_{dist} ($p < .01$). In this category, only $\text{BOW}^{\text{dim-red.}}$ beats the baselines ($p < .001$). Comparing the top structured models ($\mathcal{M}_{\text{wordsim}}$) with the baselines, $\text{BOW}^{\text{dim-red.}}$ and the best model from the literature Lin_{dist} we find that the ability of $\text{BOW}^{\text{dim-red.}}$ to distinguish the correct candidate from the distractors outperforms all others ($p < .001$).

Model	acc	acc_{cov}	cov
Baselines and word-based DSMs			
Random	.25	.25	**1**
Frequency	.31	.31	**1**
BOW	.46	.46	.98
$\text{BOW}^{\text{dim-red.}}$	**.55**	.55	.98
Syntax-based DSMs – $\mathcal{M}_{\text{wordsim}}$			
DM_{DE}	.48	.53	.84
$\text{DM}_{\text{EN}\rightarrow\text{DE}}$	.46	.61	.58
Models from the literature – [MGHZ07]			
Lin_{dist}	NA	.52	.45
HPG	NA	**.77**	.22
RPG	NA	.69	.27
JC	NA	.44	.36

Table 6.5.: Task 2: Accuracy and coverage for synonym choice on the Reader's Digest Word Choice dataset. MGHZ07: Mohammad et al. (2007). (Highest values in each column are in boldface.)

While the W×LW wordsim model for $\text{DM}_{\text{EN}\rightarrow\text{DE}}$ beats the best literature model and the random baseline ($p < .001$), it does not significantly outperform the frequency baseline. The nominally higher accuracy is rendered insignificant by its lower coverage. By contrast, the monolingual DM_{DE} as W×LW perform significantly better than all other models and baselines with the exception of $\text{BOW}^{\text{dim-red.}}$.

Leaving coverage out of consideration, Mohammad et al.'s gloss-based model (HPG) is the most precise, i.e. it has the highest covered accuracy acc_c. It seems the specific information provided by the definitions provides a high level of quality for detecting synonyms.[11] Hierarchical GermaNet models (Lin_{dist}, JC) do show a higher coverage than the gloss-based ones, but also markedly lower overall accuracy around .5. The (S)DSMs show the highest coverage of all models considered at 84% for DM_{DE} and 58% for $\text{DM}_{\text{EN}\rightarrow\text{DE}}$ and near full coverage for the BOW spaces. Our DM-based models are purely distributional, however, and

[11] However, this information is not always available; leading to a low coverage with only around one in four items is covered.

do not require any manually constructed, large-scale semantic knowledge base allowing distributional models to achieve superior coverage.

We also took a closer look at the DM model with the lowest acc_{cov}, finding a difference between the word and phrase models. The latter reliably outperforms the former: .55 acc_{cov} for the word model versus .50 for phrase ($p < .001$). And this comparison holds even when measured on the entire dataset with each model abstaining on non-word or non-phrase items, respectively. This is somewhat surprising as the phrase model does no further analysis to determine phrase structure or meaning. Instead, its simplistic heuristic compares the target word to each single word within the phrases which can be assumed to contain less information than a single word item. A closer look at the phrasal candidates reveals why our models do better on these items. The example in Table 6.6 shows, they often resemble defining phrases or glosses.

As a consequence, on each single phrasal candidate, our W×LW models are in principle attempting to solve a simplified Lesk-style disambiguation task, similar to Mohammad et al.'s gloss-based models before the maximal similarity across candidates is used for predicting the synonymous candidate, If we consider the quantifying of shared contexts in the vector space to correspond – albeit loosely – to the overlap in the glosses in the disambiguation process. As an example, the common word *Witz*, while not an exact synonym of the target word *Kalauer*, would most likely be used when defining it. At the same time, the single-word candidates handled by the word model are often more rare (e.g. *Thanatologie*) or words whose intended meaning (in the task) is very much a minority sense (e.g. *Amerikaner* in Table 6.6). The performance difference between our two models remains fairly small overall. The non-zero ties arise in phrasal items that include the same phrase with subtle variations, e.g. *Sommertag* (summer day) as the correct candidate has the candidate phrases: [*Höchsttemperatur*] *von mindestens* $25°/20°/28°/30°$ *C* (high temperature of $77°/68°/82.4°/86°$ F). This is obviously a limitation of the simple heuristic for handling phrases which fails to distinguish between candidates that share the most similar word.

target	Correct cand.	Distractor candidates		
Kalauer old chestnut	*billiger [Witz]* cheap joke	*leichte [Kutsche]* light carriage	*steifer [Hut]* stiff hat	*[Merkspruch]* mnemonic
Caddie golf trolley	*Golfwägelchen* golf trolley	*Golfschläger* golf club	*Golfplatz* golf course	*Golf-Abschlag* tee off
Thanatologie thanatology	*Sterbekunde* study of death	*Traumdeutung* dream interpretation	*Erblehre* heredity studies	*Zukunftsforschung* future studies
Amerikaner black-and-white cookie	*Feingebäck* pastry	*Cowboyhut* cowboy hat	*Wirtschaftszeitung* business newspaper	*Hochhausvariante* skyscraper style

Table 6.6.: Example single-word and phrase items. The bracketed words in the phrasal items are those with the highest similarity to the target in DM_{DE} wordsim model.

6.3. Task 3 – Argument Plausibility

Reminder of task. In this task we increase the syntactico-semantic complexity according to Figure 3.5: Here, we investigate the plausibility of predicate-argument combinations across three relations types – subject, object and prepositional object.

Evaluation measures. Coverage and Pearson's correlation coefficient r between model predictions and human plausibility judgment is reported.

DM models. W×LW models are not well-suited to this task as we are in every case comparing across parts of speech. As links are encoded in the vector space, most similarities are low – instead, it is best to perform vector similarities to words of the same part of speech. The first two models attempt to determine a realistic vector for the prototypical slot filler for a given relation.

$\mathcal{M}^k_{head}$: In this model the prediction is the similarity of the dependent's W×LW vector to the prototype based on the top k dependents related via the link REL to head:

$$\pi_{\mathcal{M}^k_{head}}(\langle head \text{ REL } dep\rangle) = cos\,sim\left(\sum_{d \in \text{deps}^k_{\text{REL}}(head)} \vec{d},\ \overrightarrow{dep}\right)$$

$\mathcal{M}^k_{dep}$: Same as above, only this time we calculate the similarity of the head's vector to the dependent-based prototype:

$$\pi_{\mathcal{M}^k_{dep}}(\langle head \text{ REL } dep\rangle) = cos\,sim\left(\sum_{h \in \text{heads}^k_{\text{REL}}(dep)} \vec{h},\ \overrightarrow{head}\right)$$

Example. In order to compute the plausibility for the item ⟨*Pfarrer* SUBJECT *beten*⟩ (⟨*pastor* SUBJECT *pray*⟩) according to $\mathcal{M}_{head}$, we first determine the top-k dependents for the syntactic head of the combination, i.e. the predicate $beten$:

$$\begin{aligned}\text{deps}^{k}_{\text{SBJ\_INTR}}(\textit{beten}) &= top_k(\lambda w.\sigma(\langle \textit{beten} \text{ SBJ\_INTR } w\rangle)) \\ &= \{\textit{Mensch}, \textit{Christ}, \textit{Jesus}, \textit{Paulus}, ...\} \qquad \text{(man, Christian, Jesus, Paul)}\end{aligned}$$

$$\pi_{\mathcal{M}^{k}_{head}}(\langle \textit{beten} \text{ SUBJECT } \textit{Pfarrer}\rangle) = cos\,sim\left(\overrightarrow{\textit{Mensch}} + \overrightarrow{\textit{Christ}} + ... , \overrightarrow{\textit{Pfarrer}}\right)$$

EPP: Distributional selectional preference model Erk et al. (2010)

Similarly to our dep and head prototype models, EPP models the plausibility of a dependent for a given $\langle \textit{head} \text{ REL}\rangle$ combination using relational and distributional knowledge. The weighted similarity of the candidate dependent vector $\overrightarrow{dep}$ with those of acceptable dependents' vectors is then taken to be its plausibility for the triple $\langle \textit{head} \text{ REL } \textit{dep}\rangle$.

$$\pi_{\text{EPP}}(\langle \textit{head} \text{ REL } \textit{dep}\rangle) = \sum_{dep' \in deps_{\text{REL}}(head)} \frac{\sigma(\langle \textit{head} \text{ REL } \textit{dep'}\rangle)}{Z_{\langle \textit{head} \text{ REL } \cdot\rangle}} \cdot cos\,sim(\overrightarrow{dep'}, \overrightarrow{dep}) \quad (6.1)$$

with normalization constant $Z_{\langle \textit{head} \text{ REL } \cdot\rangle} = \sum_{dep' \in deps_{\text{REL}}(head)} \sigma(\langle \textit{head} \text{ REL } \textit{dep'}\rangle)$.

EPP can accordingly be viewed as an **exemplar model** (Nosofsky, 1986; Daelemans and Van den Bosch, 2005; Nosofsky, 2011) in that it relies on a memory store of all known dependents for each combination. The main difference as compared to our prototype models, EPP considers all arguments ever encountered and weights them according to their association score. The comparison of the vectors of a typical[12] noun argument to a verb conforms with the results of McRae, Ferretti, and Amyote (1997) showing high overlap in psychologically salient features of the verb's typical role fillers with those of the noun.[13]

W×LW word similarity models were also considered but as the dataset is designed

[12] The inclusion of a measure of the **typicality** of an argument slot filler naturally corresponds in EPP to the weighting of each similarity in Equation 6.1, with more plausible argument slot-fillers contributing more the overall prediction π_{EPP}.

[13] Erk et al. (2010) originally introduced EPP as a model of regular and inverse selectional preference of a predicate on its arguments corresponding to our grammatical link types such as subject and object. When using an SDSM as the fundamental model, any link can be used to model prototypes.

in such a way that the pairs always contain a noun and a verb, the similarity in a structured space can expected to be zero, which is what we find.[14] We thus omit DM word similarity models from this experiment.

Non-DM models.

- CondP: is the conditional probability of seeing the verb given the noun for that relation:

$$P(v|n;r) = \frac{f(v,n;r)}{f(n;r)} \tag{6.2}$$

Here, the plausibility of a verb-noun combination is modeled as the probability of seeing the verb as subcategorizing the noun. This is a more reasonable model than $P(n|v;r)$ as there are fewer verbs than nouns in our vocabularies.[15] E.g. among the grammatical relations between nouns and verbs present in DM, there are on average 85 nouns per verb and only 7 verbs per noun. It should thus be expected that the probability estimates will be more reliable as the probability mass is distributed between fewer collocants.

- SelA: the selectional association model proposed by Resnik (1993)

In this model, semantic classes are used to smooth the probability estimates for particular nouns:

$$\pi_{\text{SelA}}(v,n;r) = \frac{1}{Z} \cdot max_{c_i \in \text{classes}(n)} \left[P(c|v;r) \cdot \log \left(\frac{P(c|v;r)}{P(c)} \right) \right]$$

where Z is the normalization constant, the sum of the bracketed expression over all classes c_i. In addition to the reliance on target language corpus frequencies for estimating the probabilities, it also requires a target language taxonomy in which

[14] This is related but not identical to the situation of the overall decrease similarity due to the increased specificity in SDSMs wrt. DSMs mentioned in the previous section. For SDSMs it is one of its benefits that we lose spurious similarities (cf. the case of *hunter* vs. *deer*). In the case of cross-part of speech similarities, we can expect this decrease to be significantly more noticeable.

[15] The reason for this has been hypothesized as being due to their referents being more accessible and less conceptually complex (cf. the Natural Partitions Hypothesis (Gentner 1982; 2006)).

nouns are mapped to concepts or semantic classes (e.g. in the case of a WORDNET, synsets). The conceptual advantage of this formulation over the direct frequency model is that only the most relevant class which links the verb and noun together in a strongly associated way determines the predictions.

Baseline.

- Freq: raw corpus frequency on German Süddeutsche Zeitung of the verb-noun pair occurring in the given grammatical relation:

$$f(v, n; r)$$

The results for the above three non-DM models are listed as reported by Brockmann and Lapata (2003).

Results. Table 6.7 shows the correlation values for all models considered. While the conditional probability model and Resnik's selectional association model perform best overall when taking all relation types into account, the frequency baseline does comparatively well on subject and object. Only one of our models produced significant correlation values across parts of speech (DM_{DE}, head prototype model). This is disappointing but not surprising as 10 out of 15 original experiments (5 models, 3 data segmentations by relation) performed by Brockmann and Lapata (2003) failed to yield significant correlations. One of the reasons for this is that the number of experimental items is quite low: only 30 word pairs per relation, totaling 90.

Comparing the bare values of all significant correlation values in Table 6.7 only our DM_{DE} significantly outperforms the frequency baseline.[16]

Another effect of interest is that models can be good at modeling single relation types but fail to provide stable predictions across relations. In particular, most

[16]The coverage values for the models were not reported by Brockmann and Lapata (2003) and as such, could not be considered. The significance of the difference between correlation values was computed using a one-tailed Fisher's z-test with Bonferroni correction.

Model	Relation			
	Subj.	Obj.	PPobj.	All
Baseline & Literature models (Brockmann and Lapata, 2003)				
Freq	.386 (*)	.360	.168	.301 (**)
CondP	.010	.399 (*)	.335	**.374** (***)
SelA	.408	.430	.330	**.374** (***)
First-order models – $\mathcal{M}_{\text{highest}}$				
DM_{DE}	.160	.310	-.299	.153
$\text{DM}_{\text{EN}\rightarrow\text{DE}}$	.254	-.157	.042	.002
First-order models – $\mathcal{M}_{\text{gram}}$				
DM_{DE}	.187	.267	-.299	.153
$\text{DM}_{\text{EN}\rightarrow\text{DE}}$	-.087	.202	.214	.089
Prototype models – $\mathcal{M}^k_{\text{head}}$				
DM_{DE}	**.528** (k=10, **)	**.478** (k=20, **)	.172 (k=1)	.309 (k=20, **)
$\text{DM}_{\text{EN}\rightarrow\text{DE}}$	.141 (k=100)	.250 (k=10)	**.413** ($k = 2$, *)	.148 (k=5)
Prototype models – $\mathcal{M}^k_{\text{dep}}$				
DM_{DE}	.325 (k=1)	.345 (k=100)	.367 ($k = 20$, *)	.194 ($k = 100$)
$\text{DM}_{\text{EN}\rightarrow\text{DE}}$	.122 (k=2)	.331 (k=2)	.181 ($k = 1$)	.228 ($k = 1$)
EPP models (Erk et al., 2010)				
DM_{DE}	.215 (†: reg)	.187 (†: inv)	.089 (†: reg)	.183 (†: reg/inv/reg)
$\text{DM}_{\text{EN}\rightarrow\text{DE}}$	.204 (†: reg)	.231 (†: inv)	-.038 (†: reg)	.044 (†: reg)

Table 6.7.: Task 3: Correlation values (r) on Brockmann and Lapata (2003) dataset. We report correlation on covered items. For EPP models the dagger (†) indicates the directionality of the similarity calculations: 'reg' refers to the regular selectional restrictions in which the given dependent is compared to all other seen dependents; 'inv' does the reverse, comparing the head vector to all seen heads' vectors. For any given relation type, we choose the directionality which to maximal correlation. Cases in which the correlation cannot be computed – due to lack of coverage or singularities stemming from zero predictions – are denoted as '-'. (Highest values in each column are in boldface.)

models have a difficulty modeling the class of prepositional objects with our cross-lingual first-order models hurting most from coverage issues or low estimates.[17]

[17]Caution must be exercised when dealing with correlations with many 0 predictions,

This can be seen both in our best model, DM_{DE}, and a number of models investigated by Brockmann and Lapata (2003). In summation, our prototype models can model the single grammatical relations better than models from the literature that have explicit semantic classes from manual annotations (SelA).

We conclude that k=20 is a robust setting for prototype construction as it was the only parameter value among those tried that led to multiple significant correlations, even across models.

6.4. Task 4 – Logical Metonymy

Reminder of task. In terms of compositional complexity (cf. Figure 3.5), this task is at the peak: Now we are faced with making predictions for combinations of multiple arguments. Moving beyond mere subject or object combinations the basis of this task are $\langle s, v, o, e\rangle$ tuples which in contrast to the previous tasks, will highlight the ability of our models to integrate multiple sources of contextual semantic information in one prediction. In dealing with this integration of multiple elements, this task is firmly in the setting of the theory of semantic compositionality.[18] Due to the complex nature of this topic within linguistic theory the goal of developing a fully compositional semantic model lies well beyond the scope of this thesis. Instead, we content ourselves with addressing the ability of our SDSMs to reliably integrate the single points of composition – without exhaustively plotting out the search space for a general theory-based model of semantic composition.

The design of the psycholinguistic experiment that yielded this dataset was such that each subject-object pair $\langle s, o\rangle$ occurs once each in a high- and in a low-thematic

as the Pearson coefficient is highly sensitive to outliers; an investigation showed that using the plausibility judgments in the present dataset, randomly assigning a non-zero value to any one item has a chance of approximately 10% of leading to a significant correlation.

[18] In linguistics, this has traditionally meant the **principle of compositionality** regarding the derivation of linguistic meaning (Frege, 1884) from syntactic structure. Montague (1970) represents the first concrete and extensive algebraic formulation of the principle, which stipulates a strict constraint of the existence of a homomorphism – i.e. a total conservation of structure – between syntax trees and their corresponding semantic operations. We make no claim to be able to approach the issue on this level here, as this topic alone has in the last decade warranted entire theses in order to be approached (Clarke, 2007; Kartsaklis, 2015).

fit condition with covert events e_i ($i \in 1, 2$), with the metonymic verb as the foil for triggering the generation of a metonymic interpretation. Among the models described below, both probabilistic models – which have a natural notion of compositionality via the chain rule for joint distributions – as well as our SDSMs which allow for the use of the link structure provide a straightforward method for composition.

Evaluation measures. Due to the dataset design, from the original 96 tuples, only the 48 pairings of corresponding high- and low-thematic fit combinations $\langle s, v, o, e_i \rangle$ ($i \in \{1, 2\}$) can be assessed using accuracy and coverage.

DM models.

$\mathcal{M}_{\text{highest}}/\mathcal{M}_{\text{gram}}$: First-order models as introduced above

Similar to the experiments in the previous task, we investigate the extent to which the links present in our mono- and cross-lingual SDSMs are themselves able to cover and directly capture the syntactico-semantic relationship at hand. However, since we have multiple constituents on the basis of whose relationship with the covert event we can make predictions, we further split the highest and grammatical classes up into three models: s, o and v. For the s gram models we use the subject in the given sentence and only extract the first-order weights of the intransitive subject edge between the subject and the covert event verb, if present. Similarly for the o model we use direct object links, and for the v model: vcomp edges. For highest models, we proceed as in prior tasks selecting the highest score independent of link type. Note that due to the balanced design of the dataset, objects and metonymic verbs co-occur equally often with the same covert event in high- and low-typicality conditions. We include the results to show the coverage of these object/verb-covert event pairs in the DMs as an indication of their utility for similar experiments.

$\mathcal{M}^k_{\text{dep/head}}$: Subject prototype models

In these models, the prototype is built with respect to the relationship between subject and covert event. As the subject of the matrix, i.e. outer, clause is also the implicit subject of the event verb we can investigate SUBJ_TR and SUBJ_TR^{-1}

prototypes coming from the verb and subject, respectively. Following the results in the previous task, we will set the parameter k to 20.

ECU models: Expectation Composition and Update model (ECU: Lenci, 2011), a natural extension to the prototype models introduced in the preceding section.

Among the approaches to modeling compositionality using DSMs in a linear algebraic framework there are two extremes: The first and most simple form proposed by Mitchell and Lapata (2010) was component-wise operations – in which any syntactic information is lost. On the other hand, there are models that fully capture the argument-function type relationships between constituent but that incur additional computationally expensive training, use and storage of the required lexicalized matrix or tensor representations (Baroni and Zamparelli, 2010; Guevara, 2010; Wu and Schuler, 2011; Kartsaklis and Sadrzadeh, 2016); although there has been promising work investigating the quality of lower-rank tensors approximations in modeling verb compositionality (Fried, Polajnar, and Clark, 2015).

The basis for the ECU approach to modeling the acceptability of verb-argument combinations lies in the results of psycholinguistic studies showing that presenting readers with verbs activates expectations on nouns and vice versa (McRae, Spivey-Knowlton, and Tanenhaus, 1998; McRae, Hare, Elman, and Ferretti, 2005). In addition to single-link plausibilities, e.g. as seen in EPP (cf. Equation 6.1), ECU has a notion of the composition of the association between verbs and their arguments that is encoded in terms of expectations that are updated with increasing contextual information. In a final step, similar to the $\mathcal{M}^k_{\text{head}}$ and $\mathcal{M}^k_{\text{dep}}$ models described in the last task, ECU compares the word W×LW-vector to the prototype defined as the centroid of the vectors of the top k words ranked by their composed expectations.

In concrete terms, the multi-step calculation of expectations for an object in a subject-verb-object triple ECU uses the DM scores coming from both the verb:

$$EX_V(v) = \lambda o.\sigma(\langle v \text{ OBJ } o \rangle)$$

and the subject:

$$EX_S(s) = \lambda o.\sigma(\langle s \text{ VERB } o\rangle)$$

to finally compute a combined object expectation:

$$EX_{SV}(s,v) = \lambda o.EX_V(v)(o) \circ EX_S(s)(o)$$

where $\circ$ is a composition operation which in Lenci's original experiments instantiates to component-wise sum and product (cf. Mitchell and Lapata (2010)).

The locus of plausibility – or typicality – judgments in the case of our $\langle s, v, o, e\rangle$ tuples is the covert event *e* and there are thus three impinging context expectations available to include in our models:

$$EX_S(s) = \lambda e.\sigma(\langle s \text{ SUBJ } e\rangle)$$
$$EX_O(o) = \lambda e.\sigma(\langle o \text{ OBJ } e\rangle)$$
$$EX_V(v) = \lambda e.\sigma(\langle v \text{ COMP } e\rangle)$$

which would yield a different model for each of the 7 non-empty subsets of:

$$\{EX_S(s), EX_O(o), EX_V(v)\}.^{19}$$

For example, for an item containing a tuple such as ⟨*Chauffeur vermeiden Auto reparieren*⟩ (⟨chauffeur avoid car repair⟩) this means we can combine the expectations from any combination of the subject *Chauffeur*, the metonymic verb *vermeiden* as well as the object *Auto* to assess the typicality of the event *reparieren*. To simplify the experiment, we build models analogously to the probabilistic models constructed by Lapata, Keller, and Scheepers (2003) which are described below.

> $SO_{\Sigma/\Pi}$: Subject-object ECU models with combination functions $+$ (Σ) and $\cdot$ (Π)

[19] We assume commutativity of the underlying composition function which means every non-empty subset leads to an ECU model for our data, e.g. $\{EX_O(o), EX_V(v)\} \mapsto \mathcal{M}_{EX_{OV}}$ and $\pi_{\mathcal{M}_{EX_{OV}}} \equiv \pi_{\mathcal{M}_{EX_{VO}}}$.

$$EX_{SO}(s,o) = \lambda e.EX_S(s)(e) \circ EX_O(o)(e)$$

$SOV_{\Sigma/\Pi}$: Subject-object-verb ECU models (combination functions as above)

$$EX_{SOV}(s,o,v) = \lambda e.EX_{SO}(s,o)(e) \circ EX_V(v)(e)$$

In terms of the two standard choices for the composition function, multiplication (in Π models) can be viewed as a form of conjunction, promoting objects that are strongly preferred by subject and object, or subject, object and verb. This means, however, that it will be prone to sparsity problems compacting the issue of sparsity of the underlying SDSM. The sum composition function (Σ models) acts more like a disjunction in that it suffices for an object to be highly expected via any of the subject, object or verb's contribution but not necessarily all.

For example:

$$\pi_{SO_{\Sigma/\Pi}}(\textit{Mechaniker vermeiden Auto reparieren}, \textit{Chauffeur vermeiden Auto reparieren}) = \chi\,(\alpha > \beta)$$

where:

$$\alpha = \textit{cos sim}\left(\overrightarrow{\textit{reparieren}}, \text{prototype}(\textit{Chauffeur}, \textit{Auto}, SO)\right)$$

$$\beta = \textit{cos sim}\left(\overrightarrow{\textit{reparieren}}, \text{prototype}(\textit{Mechaniker}, \textit{Auto}, SO)\right)$$

where the prototype vector is defined similar to the dep/head models in Task 3:

$$\text{prototype}(\textit{Mechaniker}, \textit{Auto}, SO) = \sum_{e \in \text{top}_k(EX_{SO}(\textit{Mechaniker},\textit{Auto}))} \vec{e}$$

where e top$_k$ select those top k words according to the expectations, i.e. have the highest combined expectations coming from both the subject *Mechaniker* and the object *Auto*. Such component-wise operations $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ are the most frequently used **vector mixture models** which despite being very simple prove surprisingly effective and hard to beat in compositional distributional semantics tasks (Blacoe and Lapata, 2012; Kartsaklis, 2015).

We will refer to this model as SOV_Σ when the composition function is sum, and as

the SOV_{Π} model when the composition function is product. It would be possible to use the composed expectation scores of ECU as direct estimates of expectations, i.e. to read off for a specific object its value, however, to overcome sparsity, ECU opts to compare a given object's vector with the prototype of the top 20 expected objects according to EX_{SV}, thus the ECU prediction for an item $\langle s, v, o, e\rangle$ becomes:

$$\pi_{SOV_{\Sigma/\Pi}}(\langle s, v, o, e\rangle) = cos\, sim(\vec{o}, \sum_{o' \in top_{20} EX_{SOV}(s,o)} \vec{o'})$$

Non-DM models. Prior work on logical metonymy interpretation focused on probabilistic models (Lapata et al., 2003; Lapata and Lascarides, 2003) which view the construction of the sentence as a random process assigning values to the slots of interest. On this view, the acceptability of a phrase $\langle s, v, o, e\rangle$ (e.g. *der Author begann das Buch zu lesen*, the author began to read the book) is the probability $P(s, v, o, e)$ over s (the subject, *Author*), v (the metonymic verb, *beginnen*), o (the object, *Buch*) and e (the covert event, *lesen*). This is similar to the CondP model of Task 3 which modeled simple argument plausibilities as their conditional probability (cf. Equation 6.2). The components of a given phrase are then random variables whose joint probability distribution can be modeled in different ways by making varying assumptions of independence. The advantage in using this formulation is that it provides a straightforward way to include or exclude context, by simply adding or leaving out random variables from the probabilistic model. These assumptions may or may not be appropriate, and the models originally investigated by Lapata et al. only take first-order co-occurrence evidence into account, compared to the similarity-based second-order associations of the selectional preference models of EPP and ECU.[20]

SOV_p: Probabilistic model using subject, object and verb

Lapata et al. develop a model which we will refer to as the SOV_p model.[21] It

[20] This can be remedied with more complex probabilistic models, e.g. employing generative models that introduce latent variables. This would be analogous to clustering based on higher-order co-occurrences, (cf. e.g. Prescher, Riezler, and Rooth, 2000).

[21] In Lapata et al. (2003); Lapata and Lascarides (2003), this model is called the simplified model to distinguish it from a fully specified model. Since the full model performs worse, we do not include it into consideration and use a more neutral name for the simplified model.

assumes a generative process which first generates the covert event e and then generates all other variables based on the choice of e:

$$\pi_{SOV_p}(\langle s, v, o, e\rangle) = P(s, v, o, e) = P(e) \cdot P(o|e) \cdot P(v|e) \cdot P(s|e)$$

The factorized distributions are estimated in the standard way using maximum likelihood estimation on SDEWAC:

$$\begin{aligned}
\hat{P}(e) &= \frac{f(e)}{N} \\
\hat{P}(o|e) &= \frac{f(o, e)}{f(\cdot \overset{o}{\leftarrow} e)} \\
\hat{P}(v|e) &= \frac{f(v, e)}{f(\cdot \overset{v}{\leftarrow} e)} \\
\hat{P}(s|e) &= \frac{f(s, e)}{f(\cdot \overset{s}{\leftarrow} e)}
\end{aligned}$$

where N is the number of occurrences of full verbs in the corpus; $f(e)$ is the frequency of the verb e; $f(\cdot \overset{o}{\leftarrow} e)$ and $f(\cdot \overset{s}{\leftarrow} e)$ are the frequencies of e with a direct object and subject, respectively; and $f(\cdot \overset{v}{\leftarrow} e)$ is number of times e is the complement of another full verb.

SO_p: Probabilistic using subject and object

In the covert event data used by Lapata et al. (2003); Lapata and Lascarides (2003), v, the metonymic verb, was used to prime different choices of e for the same object. In our dataset (Sec. 4), we keep v constant and consider e only as a function of s and o. Thus, in the second model we do not consider v:

$$\pi_{SO_p}(\langle s, v, o, e\rangle) = P(s, o, e) = P(e) \cdot P(o|e) \cdot P(s|e)$$

estimated as above.

Baselines. Due to the balanced nature of the dataset, the design of a frequency baseline that only compares the covert event verb with with the object or the metonymic verb can only give a 50% accuracy, leaving only the subject as a

Model	acc	acc_{cov}	cov
Baselines			
Random	0.50	0.50	1.00
Freq_s^Δ	**0.63**	0.63	1.00
$\text{Freq}_{s,v}^\Delta$	0.60	0.60	1.00
Probabilistic Models			
S_p	0.54	0.57	0.96
SO_p	0.56	**0.75**	0.75
SOV_p	0.27	0.62	0.44

Table 6.8.: Performance of baseline and probabilistic models on Task 4.

potential identifier. Experimenting with integrating the object and metonymic verb frequencies, we indeed find that accuracy decreases when adding either (or both) of these to the subject. We report the top two frequency baselines, Freq_s^Δ and $\text{Freq}_{s,o}^\Delta$, which predict that $\langle s, v, o, e\rangle$ tuple as exhibiting higher typicality for which the frequency of e is closer to s, respectively, the mean of s and o.

Results. Table 6.9a shows the performance of our DM edges models on the logical metonymy dataset. As previously mentioned, all single o/v-type models have a covered accuracy of 50% due to the balanced dataset. The differences in total average scores comes from the differences in coverages of the models. Coverage varies between monolingual and cross-lingual DMs and model types: The more restrictive gram models obviously cover fewer items than the highest ones. Except in the case of $\text{DM}_{\text{EN}\to\text{DE}}$ highest subject model, corresponding cross-lingual models always have less or equal coverage to the monolingual models; the higher coverage (compared to the DM_{DE} highest subject model) comes at the cost of accuracy (0.71 compared to 0.92). Among the subject models, the top performing is the $\text{DM}_{\text{EN}\to\text{DE}}$ highest first-order model which significantly outperforms the gram models $\text{DM}_{\text{EN}\to\text{DE}}$ ($p < .01$) and DM_{DE} ($p < .05$).

Unsurprisingly, we find among the probabilistic models a noticeable difference in terms of coverage between the model with more parameters (SOV_p: .75) and with fewer parameters (SO_p: .44). SOV_p is unable to make predictions on over

Model	Type	acc	acc_{cov}	cov
Highest first-order Models – $\mathcal{M}_{\text{highest}}$				
DM_{DE}	s	0.25	0.92	0.27
	o	0.42	(0.50)	0.83
	v	0.19	(0.50)	0.38
$\text{DM}_{\text{EN}\to\text{DE}}$	s	0.31	0.71	0.44
	o	0.42	(0.50)	0.83
Grammatical first-order Models – $\mathcal{M}_{\text{gram}}$				
DM_{DE}	s	0.06	**1.00**	0.06
	o	0.42	(0.50)	0.83
	v	0.06	(0.50)	0.13
$\text{DM}_{\text{EN}\to\text{DE}}$	s	0.04	0.67	0.06
	o	0.19	(0.50)	0.38

(a) First-order models. Here, single edge weights are used to determine the more typical covert event.

Model	Type	acc	acc_{cov}	cov
s Prototype Models – $\mathcal{M}^k_{\text{dep/head}}$				
DM_{DE}	dep	0.52	0.53	0.98
	head	0.48	0.50	0.96
$\text{DM}_{\text{EN}\to\text{DE}}$	dep	0.48	0.62	0.77
	head	0.52	**0.68**	0.77
ECU Models				
DM_{DE}	SO_Σ	0.67	**0.68**	0.98
	SO_Π	**0.69**	0.70	0.98
	SOV_Σ	0.67	**0.68**	0.98
	SOV_Π	0.53	0.56	0.94
$\text{DM}_{\text{EN}\to\text{DE}}$	SO_Σ	0.52	0.58	0.90
	SO_Π	0.54	0.60	0.90

(b) Vector-based models. Here, the covert event vector is compared against a prototype vector. (Parameter $k = 20$.)

Table 6.9.: Task 4: Modeling the logical metonymy dataset DM first-order models. Predictions and expectations from v are not available for $\text{DM}_{\text{EN}\to\text{DE}}$ as the source-language DM_{EN} does not cover verb-verb links. (Bracketed results are those for o and v which are included to illustrate the coverages; their accuracy is 0.50 by construction.)

half the dataset due to unattested $\langle o\, v \rangle$ combinations in the corpus. At the same time, the predictions of the SO_p model are more reliable than SOV_p (.75 vs. .62). We conclude that, at least in the case of this particular dataset, the metonymic verb does not provide useful information regarding the covert event; instead it introduces noise which overpowers the signal contributed by the subject and object. This makes sense in terms of how the dataset was designed, as no special consideration was given to the choice of metonymic verb in pairing it with covert events.

The distributional vector-based models all have significantly higher coverage scores than first-order and full probabilistic models. The items which prove problematic for our DM-based similarity models are those containing the $\langle s, v, o \rangle$ triple $\langle$*Pizzabote* HASSEN *Pizza*$\rangle$ (Pizza delivery man HATE pizza) for which the high- and low-typicality covert events are *liefern* (deliver) and *backen* (bake). In order to compute similarity predictions for this combination we require corpus instances of transitive uses for *Pizzabote*. However, in the SDEWAC corpus it is only attested once as the subject in an intransitive construction with the verb *kommen* (come).

Among distributional models, the difference between *SO* and *SOV* is not as clear-cut as on the probabilistic side. We do see a slight effect of model modalities between including the verb (*SO* vs. *SOV*) with the choice of composition operation (Σ vs. Π). On the one hand, dropping the metonymic verb in the product models helps prediction quality, similarly to the probabilistic models. This makes sense as the product models are more sensitive to local sources of noise in the chain of expectations. On the other hand, the sum models are unaffected with the contribution from the verb being overall negligible.

Comparing ECU to the probability models we can say that both in coverage and accuracy ECU does better. At the same time, the results for these two model classes seem to indicate the presence of a 'sweet spot': Going from simple subject-information to adding object increases performance while adding the metonymic verb leads to a decrease. Additionally, we also note a pattern among the s prototype models: ML DM does better than random when predictions are dep-based (i.e. when constructing a head/verb-prototype) and vice versa for XL DM. Overall, these simple prototype models perform very poorly. However the ECU results show more context can benefit prototype models.

6.5. Task 5 – Evaluation on a lower-resource language – Croatian

The goal of this thesis has been to develop methods for inducing high quality, robust structured distributional semantic models for novel languages. As these typically require significant amounts of accurately parsed data in the target language we have reduced those resource requirements to make obtaining SDSMs a realizable goal.

To show how this approach works across the resource gradient, we also evaluate our models on the language pair English-Croatian. Croatian being a Slavic language, this pairing represents a more distant relationship than English-German. These languages taken together exemplify the variability on the resource gradient: The resource situation is best for English, still relatively good for German, and more limited for Croatian.

Monolingual Croatian DM. Šnajder, Padó, and Agić (2013) made available the first SDSM for Croatian, DM_{HR}, for which they parsed a 1.2B token web corpus using a parser trained on only 4K sentences and improved on the state of the art in dependency parsing for Croatian.

Staying in line with the investigations for German, we compare Šnajder et al.'s monolingual DM_{HR} to a cross-lingual one, and are interested in seeing whether a similar pattern to the one established for German appear in their respective performance profiles.

Croatian corpus. The corpus from which both the BOW and DM_{HR} models were built is the fHRWAC (Šnajder et al., 2013). This is a filtered version of the Croatian web corpus HRWAC (Ljubešić and Erjavec, 2011) with the following modifications:

- Removal of forum or blog content – these documents contain the bulk of sentences with grammaticality issues.
- Removal of short websites as well as those that appeared to be non-Croatian – this determination was made if they:

- contained no diacritics;
- had no overlap with a list of most frequent Croatian words;
- contained any singularly foreign-language words (allowing the distinction from e.g. Serbian); or
- contained too many words of a commonly intermingled language (e.g. English/Slovene).

This results in a corpus of 51M sentences containing 1.2B tokens. These sentences were further pos-tagged (using HunPos, Halácsy, Kornai, and Oravecz (2007)), lemmatized (using CST, Ingason, Helgadóttir, Loftsson, and Rögnvaldsson (2008)) and dependency-parsed (using MST, McDonald, Lerman, and Pereira (2006)) yielding 2.6M lemmas. While such filtering steps reduce the amount of lexical data, the quality is dramatically increased (cf. SDEWAC in Section 3.2).

Cross-lingual Croatian DM. In Chapter 5, we described our method for pivoting the English DM into a new language using a translation lexicon. In an analogous fashion, we construct the cross-lingual Croatian SDSMs $\mathrm{DM}^{\mathrm{naive}}_{\mathrm{EN}\to\mathrm{HR}}$ and $\mathrm{DM}^{\mathrm{filter}}_{\mathrm{EN}\to\mathrm{HR}}$ using Taktika Nova's freely available English–Croatian dictionary[22] which contains approximately 100K translation pairs; considerably fewer than for English–German.

The resulting filtered Croatian $\mathrm{DM}_{\mathrm{EN}\to\mathrm{HR}}$ has 47K nodes and 315M edges, resulting in a density of $5.6 \cdot 10^{-6}$ which is one third that of $\mathrm{DM}_{\mathrm{DE}}$ while its number of edges is approximately one order of magnitude smaller than that of $\mathrm{DM}_{\mathrm{EN}\to\mathrm{DE}}$.

Comparison of $\mathrm{DM}_{\mathrm{HR}}$ and $\mathrm{DM}_{\mathrm{EN}\to\mathrm{HR}}$. As an evaluation scenario, we use the Croatian synonym choice dataset (CroSyn, Karan, Šnajder, and Dalbelo Bašić (2012)), corresponding to Task 2 described above for German. This dataset is larger than the German one, and has a balanced designed: For each part of speech (noun, verb and adjective) the dataset contains 1000 items, all of which are single word

[22] Available from `http://www.taktikanova.hr/eh/EHdownload_en.htm`. It provides approximately 10 times the number of translations of the EN-HR portion at dict.cc.

Model	acc	acc_{cov}	cov
Baselines			
Random	.25	.25	1
Frequency	.45	.45	1
DSM wordsim models			
BOW	.60	.60	1
$\text{BOW}^{\text{dim-red.}}$	**.66**	.66	1
DM wordsim models			
DM_{HR}	.65	.65	.99
$\text{DM}^{\text{naive}}_{\text{EN}\rightarrow\text{HR}}$	.43	.50	.71
$\text{DM}^{\text{filter}}_{\text{EN}\rightarrow\text{HR}}$	.58	**.71**	.71

Table 6.10.: Results for Croatian DMs on synonym choice task. Performance for $\text{BOW}^{\text{dim-red.}}$ and DM_{HR} was reported by Šnajder et al. (2013).

items. As was the case for the German dataset RDWP, each item contains one target and three distractors.

Table 6.10 shows the performance of these models on the synonym choice task.[23]

Results. We find that the $\text{BOW}^{\text{dim-red.}}$ and DM_{HR} perform nearly identically well and both are the only ones to outperform BOW. All models as well as the frequency baseline beat the random baseline ($p < .001$). BOW, $\text{BOW}^{\text{dim-red.}}$, DM_{HR} and $\text{DM}^{\text{filter}}_{\text{EN}\rightarrow\text{HR}}$ beat frequency ($p < .001$). Note that the naive cross-lingual model does not beat the frequency baseline, in fact, $\text{DM}^{\text{naive}}_{\text{EN}\rightarrow\text{HR}}$ is outperformed by BOW, $\text{BOW}^{\text{dim-red.}}$, DM_{HR} and $\text{DM}^{\text{filter}}_{\text{EN}\rightarrow\text{HR}}$. Despite its nominally better performance, $\text{DM}^{\text{filter}}_{\text{EN}\rightarrow\text{HR}}$ is outperformed by the BOW models and DM_{HR} ($p < .001$). This is due to the lower coverage for cross-lingual models. These results show clearly the increase in accuracy due to the backtranslation filtering which makes a difference of 15 percentage points.

We interpret the results for Croatian as encouraging and in line with the effect

[23]Models use the same configuration as in previous German experimental setup: DM_{HR} using all links, and the $\text{DM}_{\text{EN}\rightarrow\text{HR}}$ models using the selectional preference style of link selection (cf. Section 6.1).

seen in the lexical relatedness tasks for German: from a monolingual model we can expect higher coverage and fair performance while the filtered cross-lingual model reveals a complementary performance profile: high accuracy at the cost of coverage.

Again, as with German, the translational variance shows moderately strong correlation with corpus frequency among the words in the Croatian dictionary (Spearman's ρ = .46, $p < .001$). A closer look at the number of covered words (cf. Table 6.11) shows an interesting picture.

		% of Exp. words covered by	
POS	# Exp. words	Corpus	Dict.
noun	4,289	99.9	54.8
verb	3,275	99.8	49.2
adjective	3,338	98.4	42.3

Table 6.11.: Coverage of words in Croatian synonym choice experiment (CroSyn).

In fact, the dictionary only covers 2.6% of the nouns, 2.9% of the verbs and 1.5% of the adjectives present in the corpus. Applying methods to increase the dictionary size should thus give the cross-lingual model an edge.

6.6. Discussion

SDSMs can provide the basis of a variety of model types: First-order models, vector-similarity models through tensor matricization (e.g. W×LW) as well as direct (e.g. dep/head) and incremental prototype models (e.g. EPP, ECU). As such, they can form the basis of a number of experimental configurations and provide high accuracy or large coverage, depending on what is required. This chapter has shown that DMs perform comparably to or better than models which have higher resource requirements (cf. Tasks 1, 2). The structure present in SDSMs such as DM allow for predictions to be based on multiple combinations of information (e.g. in Tasks 3, 4). This is similar to the different structures for probabilistic models but SDSMs are less vulnerable to sparsity or coverage

issues.[24]

The overall trend we can see is that monolingual models tend to have higher coverage while the cross-lingual models have a higher accuracy or performance scores. In any case, with such a plethora of available models methods with which to combine them in such a way to leverage such differences in performance profiles will be the goal of the next chapter.

[24]Probabilistic models can be enhanced with smoothing techniques which raise the cost of model complexity and will likely be required for every restructuring of the model.

Chapter 7.

Combination Strategies for Multilingual DMs

7.1. From Single Source-Language to Multilingual DMs

In the preceding chapter, the benefits of the mono- and cross-lingually induced DM models have been shown to often be complementary. The monolingual methods often led to higher coverage as the model reflects the actual native language use according to the underlying corpus. However, if there are issues in the latter steps in the processing pipeline, e.g. in the parser, we end up with overall less reliable predictions.

Cross-lingually induced models have shown an inverse pattern: By relying on the output of an already heavily optimized English parser, we can expect high accuracy estimates. At the same time, since we use a bilingual dictionary to pivot such reliable English models into our target language, we can only cover those terms that have been captured in the lexicon. As we have previously argued (cf. Section 5.1.4) this can be viewed as an additional filter reducing noise in the model.

The above thus suggests the following approach: Combine the output of these two methods to obtain a **multilingual model** which would make use of the strengths of either input model while attenuating the effects of their respective limitations. The idea is that even with little or unreliable parsed monolingual target language

data, we can rely on the estimates based on the English DM to get an accurate – if low-coverage – model and then increase the coverage of our predictions with a monolingual model, e.g. in a backoff scheme: use the best available predictions first and fall back on less reliable models when necessary.

Moving beyond our two SDSM types, we can integrate a simple DSM, i.e. an unstructured model such as a BOW, into our model combination scheme which we saw perform well on a number of tasks in Chapter 6 and will have very high coverage. The reason for doing this is that there will always be more unparsed than parsed data. The goal is to make better models more widely available with a lower investment of time and computational effort.

In this chapter, we investigate methods for combining monolingually and cross-lingually constructed DMs, thereby combining corpus evidence from both the parsed source and the target language – both parsed and unparsed. To further make the methods as widely employable as possible, we restrict our considerations to combination methods that can be applied directly using only the DMs themselves without relying on the availability of any additional SL or TL resources. We assess the model combinations on our tasks at the lexical level (word relatedness and synonym choice), as these are the most general-purpose tasks with the largest number of test items.[1]

Approaches to combining models. Combining qualitatively distinct sources or models is an issue that has directly affected work on **multimodal semantic models**, not only in the computer vision research (Snoek, Worring, and Smeulders, 2005; Gunes and Piccardi, 2005) but more recently also in lexical semantics in attempts to combine the visual and lexical modalities into joint semantic representations (Bruni et al., 2014; Kiela and Clark, 2017). In the setting of multimodal research, this challenge has been framed as a question of **early**, **middle** or **late fusion** of the inputs or cues from the different modalities.

As described by Kiela and Clark (2017), the earliest point to combine evidence from different sources to create a combined model is during training in a joint learning

[1]The argument plausibility and logical metonymy datasets each have less than 100 items, while word relatedness and synonym choice cover multiple hundred items. Drawing conclusions about a model's coverage in particular should thus be more reliable on the latter tasks.

setup. In our case, we start with two pre-existing models. The question then becomes: at which point in a combined model are the two sources of information mixed, at the level of the feature representations themselves (middle fusion) or when making a prediction (late fusion)?

7.2. Middle Fusion of DMs

In our case, the middle fusion of our two DM sources, i.e. the monolingual and cross-lingual DMs, would mean combining their graphs or tensors – or the vectors of the respective matricized space. Figure 7.1 graphically represents one possible middle fusion setup. This approach of merging the graphs can be viewed as the concatenation of the W×LW matricizations of the tensors.

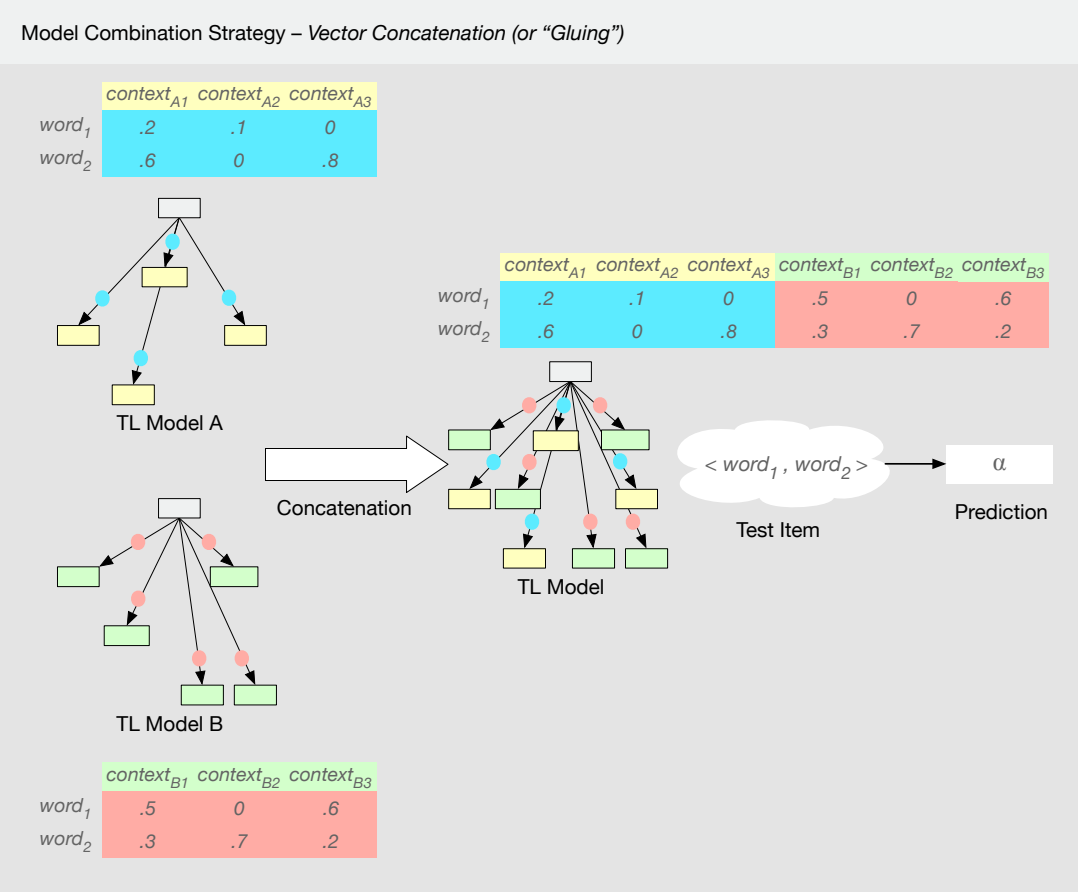


Figure 7.1.: It can be argued that concatenation will perform similarly to interpolation on a word relatedness test and be comparable to a max combination on in an ECU or highest/direct score

With such merging schemes, however, a number of questions arise regarding the details of the combination procedure.

Issues with middle fusion. Conceptually, such a procedure is making the tensor or graph more dense: the two DMs are joined in terms of graphs by adding the edges together. While the nodes are from the same base vocabulary, the edges from the cross-lingual DM come from the original SL model. Explicit labeling as either ML or XL could be employed to distinguish the source of the edges.

At the same time, potentially overlapping link types might contain interesting information, ranging from known and shared grammatical edge types (these could be merged or mapped if the relationship between SL and TL syntax were known[2]), to using shared edges as a filter. In general, this would reduce coverage significantly, and a linking or mapping of link types deemed equivalent could be used. This necessarily complicates the design of the fused model.

Additionally, it is likely the case that due to a size or density imbalance between the models (as discussed in Chapters 4 & 5), an 'over-powering' effect of one model over the other would be witnessed. A possible remedy in such a case would be to equalize the sizes of the models by sampling or partitioning the larger model. The quality of the model in such an approach then would heavily depend on the sampling or partitioning scheme applied.[3]

In summary, the many heuristic decision points outlined above yield a complex design space within which any one choice is hard to justify without a large number of experiments.

Testing middle DM fusion. Preliminary investigations into middle fusion were carried out in a Bachelor's thesis work (Aina, 2014) supervised by the author which included experiments on the combining the representations of DM_{DE} and $DM_{EN \rightarrow DE}$. The task on which the merged models were evaluated was the word relatedness task (Gur350, see Task 1 in Chapter 6).

[2]This mapping could be lexically driven: cf. the discussion of the example *The boy likes to ride his bike / Der Junge fährt gern mit seinem Rad* in Section 6.1.

[3]Additionally, putting them on equal footing can be expected to perform comparably to averaging in a late-fusion model combination.

The experimental setup consisted of middle fusion – or merging – of the single models as well as a dimensionality reduction step (cf. Figure 7.2).[4] Conceptually, the dimensionality reduction procedure would recognize latent patterns across the two models' dimensions.

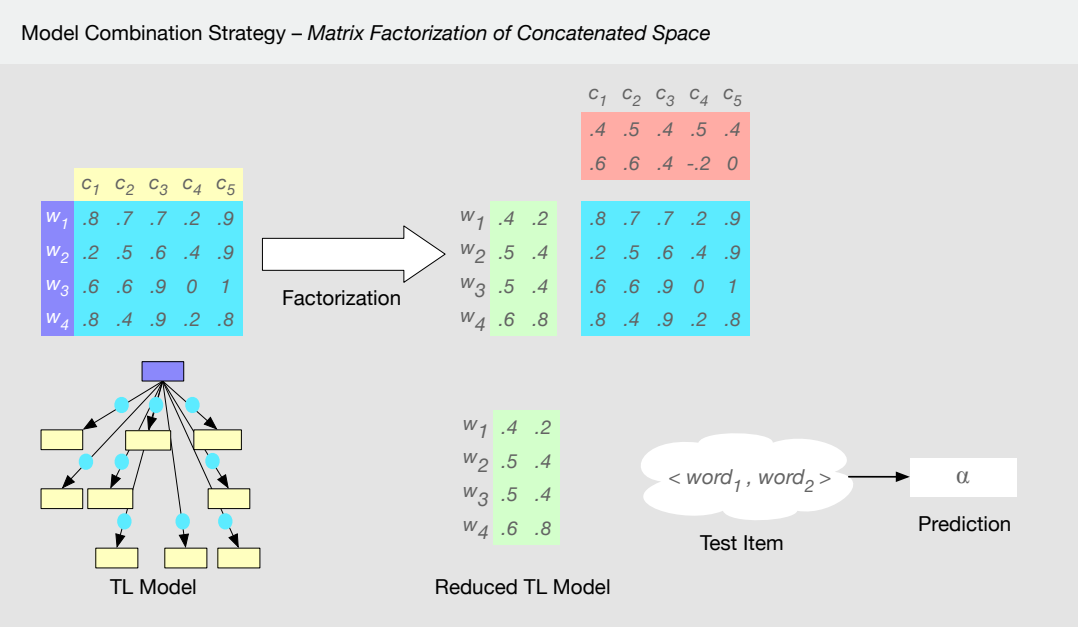


Figure 7.2.: A matrix factorization operation combines and simplifies the information in the concatenated vectors.

In order to reduce noise as well as keep the computational effort as low as possible, the W×LW spaces were also truncated to the most frequent contexts giving rise to two methods, A and B:

> Method A: truncate each W×LW space separately to 60K highest weighted LW contexts[5] then merge spaces with labels;
>
> Method B: merge first then truncate the ensuing space.

Table 7.1 shows that middle fusion at best maintains the correlation, while usually leading to a degradation in performance. In terms of the number of non-zero predictions, we do however see an improvement with dimensionality reduction. An SVD-reduced space provide somewhat more predictions while NMF leads to a

[4]We perform and compare singular value decomposition and non-negative matrix factorization to reduce the number of dimensions to 500.
[5]For each part of speech, the top 20K contexts are taken.

Model/Method	Type	r	r_{cov}	$cov.$
Single models				
DM_{DE}		**.38**	.43	.60
$DM_{EN \to DE}$		.33	**.49**	.49
Merged multilingual models				
$\mathfrak{M}^{A}_{DM_{DE}, DM_{EN \to DE}}$	full	.33	.35	.61
	SVD	.30	.29	.74
	NMF	.17	.15	**.91**
$\mathfrak{M}^{B}_{DM_{DE}, DM_{EN \to DE}}$	full	.23	.30	.36
	SVD	.24	.29	.40
	NMF	.18	.18	.50

Table 7.1.: Middle fusion results for W×LW models on the Gur350 word relatedness task. Each merging method is tested as a a full W×LW space with dimensionality reduced to 500 dimensions. Coverage captures the percentage of non-zero relatedness predictions. (Method A: truncate then merge, Method B: merge then truncate.)

clear increase. The method A approach clearly profits more however, this could be due to having twice the number of dimensions as the method B space.

A χ^2 test on the coverages reveals that all models obtained via method A have higher coverage than method B models ($p < .001$) except for full-A when compared to NMF-B. Also, both dimensionality-reduced method A models have higher coverage than the single models (NMF at $p < .001$ and SVD at $p < .05$). DM_{DE} has higher coverage than full-B and SVD-B ($p < .001$) and $DM_{EN \to DE}$ higher than full-B ($p < .05$).

Despite the nominally worse correlation values, the merged models are in general not significantly worse than the two input models: Bonferroni-corrected pairwise z-test of the correlation values of the results in Table 7.1 showed only the following significant differences:

$\mathcal{M}_1 > \mathcal{M}_2$			Sig. of comparison	
$\text{DM}_{\text{EN}\rightarrow\text{DE}}$	$>$	NMF-A	***	$(p < .001)$
DM_{DE}	$>$	NMF-A	**	$(p < .01)$
$\text{DM}_{\text{EN}\rightarrow\text{DE}}$	$>$	NMF-B	*	$(p < .05)$

where $\mathcal{M}_1 > \mathcal{M}_2$ means that $\mathcal{M}_1$ outperformed $\mathcal{M}_2$.

A final step of tensor factorization (Kolda and Bader, 2009) was considered on such a merging of the two models. However, the computational cost and complex nature of these types of approaches run counter to the spirit of this work, which is to rely on as few resources – which would include computational resources – as possible.

7.3. Late Fusion of DMs.

Whereas the previous approach attempted to join the two input models' representations with one another, we can also combine the two sources at the later stage of model predictions. The conceptual motivation for this approach can be found underpinning in the smoothing of n-gram language models (Chen and Goodman, 1999) where the operations of **summarization** or the **backing-off** from one model's predictions to another are used (Figure 7.3).

Backoff model. In a backoff model combination (cf. Figure 7.4), one would start with the model which is assumed to be more accurate and in cases in which it cannot make a prediction, fall back on the lower accuracy model. Following this principle, any number of models once given a sequential order $\mathcal{M}_1, \ldots, \mathcal{M}_N$ can be combined to form a backoff model $\mathfrak{M}^{\text{backoff}}_{\mathcal{M}_1\rightarrow\ldots\rightarrow\mathcal{M}_N}$. The higher the precedence in the ordering a model is given, the more frequently its prediction will be considered and made prevalent as the output of the combined model.

Summary model. As opposed to the backoff scheme, with a summary model all models being combined are always considered. Figure 7.4 shows how the predictions from the single models are unconditionally fed into a summary function f in order to obtain the final fused prediction. Here too any number of single models

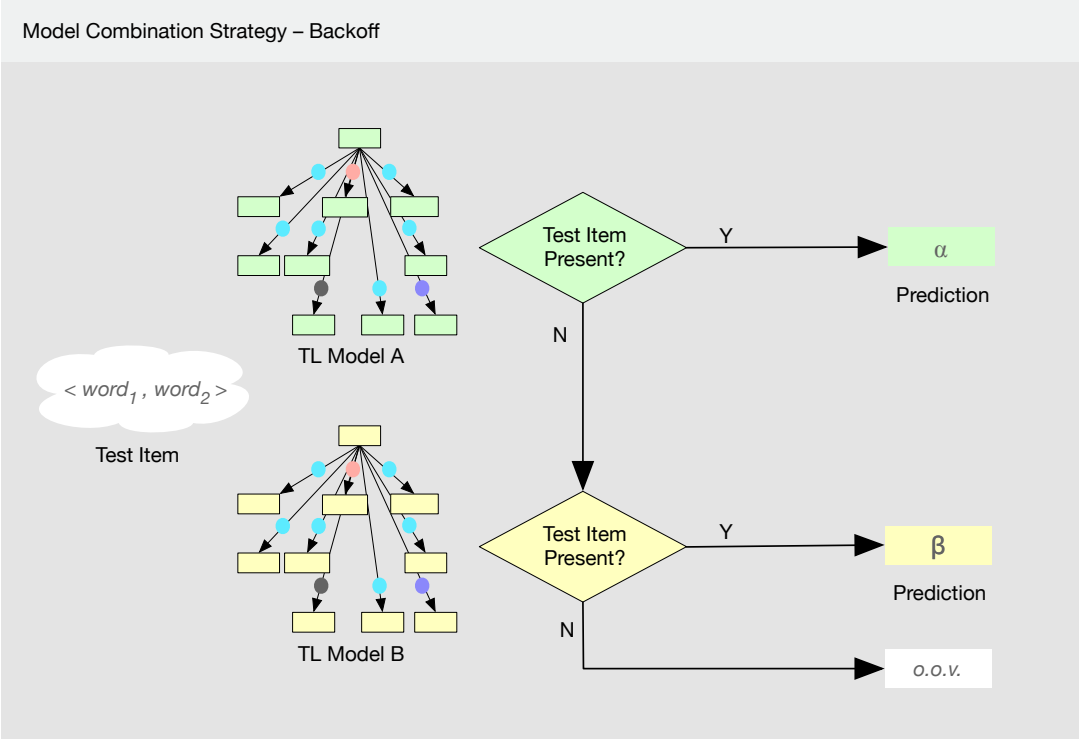


Figure 7.3.: Backing off from one model to another: In the case where the model A cannot make a prediction – or makes a prediction with low confidence – the task is handed off to model B. There is an inherent asymmetry in this scheme with the first model being the more trusted one and subsequent models decreasingly so.

can be given, depending on the arity of f. In cases where the domain of the fused model's predictions is equal to those of the single models, an *interpolation* of the single predictions is the most obvious choice:

$$\pi_{\mathfrak{M}^{\text{summary}}_{\mathcal{M}_1,\ldots,\mathcal{M}_N}} = f(\pi_{\mathcal{M}_1}, \ldots, \pi_{\mathcal{M}_N}).$$

In an effort to reduce the size of the search space for f, we decide to let each model have equal weight in the interpolation. One family of interpolation functions that is parameterization in that function space is the **generalized mean function** M_p:

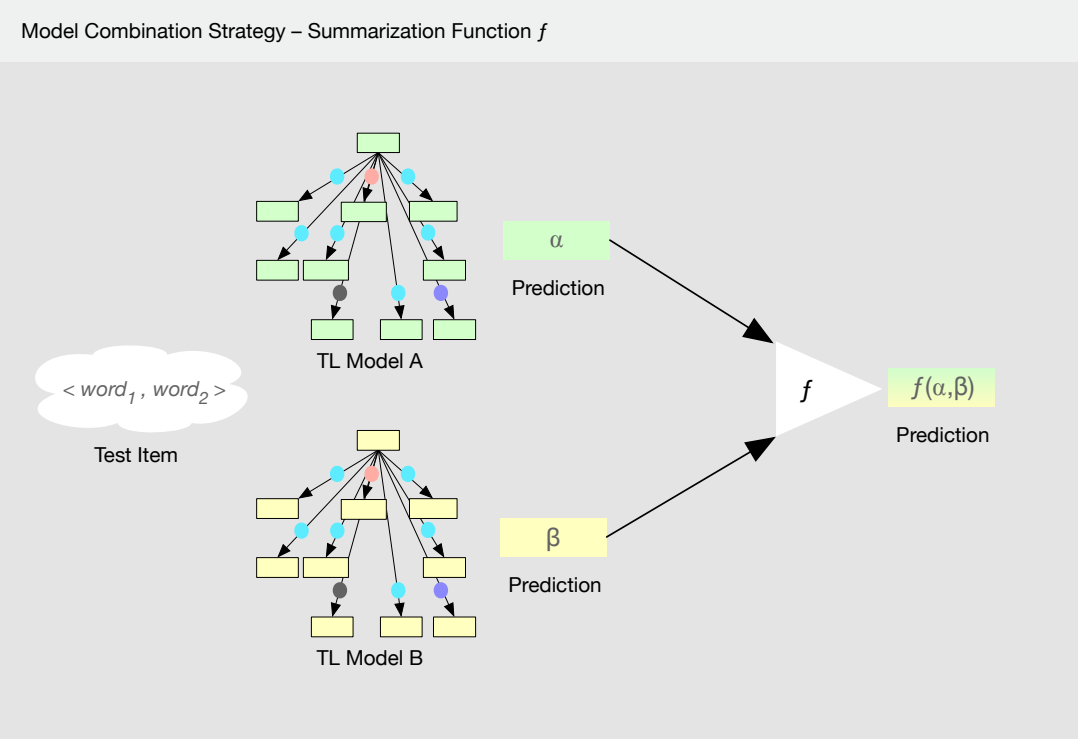


Figure 7.4.: Summarizing model predictions: Model A's predictions are directly combined with those from model B via function f.

$$M_p(x_1, \ldots, x_N) = \left(\frac{1}{N} \cdot \sum_{i=1}^{N} x_i^p \right)^{\frac{1}{p}}$$

The most commonly employed instances of the generalized mean, namely: minimum ($M_{-\infty}$), harmonic mean (M_{-1}), geometric mean (M_0), arithmetic mean (M_1) and maximum (M_{∞}).[6]

In a preliminary analysis using these five mean functions, we determined that the maximum function M_{∞} performed best, with M_{-1}, M_0, M_1 performing on equal footing and $M_{-\infty}$ worst of all. As a result, in our discussions of late fusion we will consider the maximum as a summarization function and compare it to backoff.

[6] $M_{-\infty}, M_0$ and M_{∞} are defined as the limits of M_p for $p \to -\infty, 0, \infty$, respectively. These instances obviously go beyond the simplest interpolation function, an arithmetic mean, with the extremes of $p = \pm\infty$ selecting one value out of the sequence $\{x_i\}_{i=1}^{n}$.

7.4. Evaluation of Summarization on German Data

We evaluate our models on the first two tasks presented in Chapter 3: word relatedness and synonym choice. These cover lexical relatedness in terms of both unspecified semantic relationship types as well as the specific relationship of synonymy. Following Section 6.1, we consider the selectional preference links (SPrfL) for the cross-lingual model and all links (AllL) for DM_{DE}. Our multilingual models combine DM_{DE} (AllL) with $DM_{EN \rightarrow DE}$ filtered (SPrfL).

7.4.1. Experimental Setup

As in Chapter 6, we compare bag-of-words and our DM models to monolingual ontology-based models which make use of the lexical database GermaNet, the German section of Wikipedia, or both (Lin_{GN}, HPG, JC, PL) as well as cross-lingual distributional models that represent the meaning of German lemmas using English thesaurus categories (Lin_{dist}).

While we include the BOW model in the multilingual setup as it has very high coverage and often performs reasonably well, conceivably any fair-performing off-the-shelf, high coverage model can be substituted. However since BOW models are so cheap and straightforward to generate we feel they will be the most widely available standard distributional model to use here. The full BOW space contains as dimensions 10K nouns, verbs and adjectives, with an additional dimensionality reduction step used to obtain the final 500-dimensional space.[7] As the reduced space typically outperforms the full space, we only include it in the current chapter's evaluations.[8]

[7] The dimensionality reduction method used for the BOW model was SVD to 500 dimensions. We also experimented with smaller context windows as well as building a Latent Semantic Analysis (LSA, Landauer and Dumais (1997b)) space, both with 500 dimensions and with an automatically optimized number of dimensions (Wild, Stahl, Stermsek, and Neumann, 2008). These spaces however did not consistently yield better results than the reported models.

[8] In our original papers discussing multilingual model combinations (Utt and Padó, 2014; Padó, Šnajder, Utt, and Zeller, 2016), model predictions were scaled before being fused thus leading to nominally slightly different results than those presented in this work. Qualitatively, all conclusions drawn in those studies hold regardless of scaling.

7.4.2. Late-fusion models

Backoff model. Starting with the model which consistently proved higher in accuracy, the cross-lingual DM. The backing off proceeds in linear order:

$$\mathfrak{M}^{\text{backoff}}_{\text{DM}_{\text{EN}\to\text{DE}}\to\text{DM}_{\text{DE}}\to\text{BOW}}$$

with predictions:

$$\pi_{\mathfrak{M}^{\text{backoff}}_{\mathcal{M}_1\to\cdots\to\mathcal{M}_N}}(w_1,w_2) = \begin{cases} \pi_{\mathcal{M}_1}(w_1,w_2) & \text{if} \quad N=1 \text{ or } \mathcal{M}_1 \text{ covers } w_1,w_2; \\ \pi_{\mathfrak{M}^{\text{backoff}}_{\mathcal{M}_2\to\cdots\to\mathcal{M}_N}}(w_1,w_2) & \text{otherwise;} \end{cases}$$

since $\text{DM}_{\text{EN}\to\text{DE}}$ has the highest quality, BOW the largest coverage, and DM_{DE} performs medially.

Maximum model. The predictions of a maximum model combining the models $\mathcal{M}_{1\cdots N}$ are:

$$\pi_{\mathfrak{M}^{\text{max}}_{\mathcal{M}_1>\cdots>\mathcal{M}_N}}(w_1,w_2) = max_{i\in\{1,\ldots,N\}}\pi_{\mathcal{M}_i}(w_1,w_2)$$

The full maximum model combines the estimates from the cross- and monolingual SDSMs with the unstructured DSM:

$$\mathfrak{M}^{\text{max}}_{\text{DM}_{\text{EN}\to\text{DE}}>\text{DM}_{\text{DE}}>\text{BOW}}$$

We also compare against a simpler combined model:

$$\mathfrak{M}^{\text{max}}_{\text{DM}_{\text{EN}\to\text{DE}}>\text{DM}_{\text{DE}}}$$

The combination of maximum scores is order invariant.

$\mathcal{D}(dringend_j)$	$\{Andrang_n, Bedrängnis_n, gedrängt_j, vordrängeln_v, ...\}$
$\mathcal{D}(urgent_j)$	$\{rush_n, besetment_n, terse_j, cut\ in\ line_v, ...\}$

Table 7.2.: Example of a derivational class in the German derivational lexicon DERIVBASE (Zeller et al., 2013).

7.4.3. Derivational smoothing of models

In addition, we also compare against a model, $\text{DM}_{\text{DE}}^{\text{deriv}}$ (Padó, Šnajder, and Zeller, 2013), that employs a **derivational smoothing** setup. To do this, it uses equivalence classes $\mathcal{D}$(w) of the family of words derivationally related to w in the model (cf. Table 7.2). The goal of creating such a lexicon of semantic highly related words is exactly to alleviate the effect of the fact that words that are underrepresented in the corpus should be able to make use of knowledge about their derivationally related words. Word relatedness predictions can then be written as summarizations of the similarities of word pairs across all members of these derivational classes.

In the max setup, we fully trust the maximal relatedness estimate between any two words in the two derivation classes:

$$\pi_{\text{DM}_{\text{DE}}^{\text{max}}\text{deriv}}(w_1, w_2) = max_{(w_1', w_2') \in \mathcal{D}(w_1) \times \mathcal{D}(w_2)} sim(\overrightarrow{w_1'}, \overrightarrow{w_2'})$$

In the average setup, the arithmetic mean of all word pair combinations of derivational variants is used:

$$\pi_{\text{DM}_{\text{DE}}^{\text{avg}}\text{deriv}}(w_1, w_2) = \frac{1}{|\mathcal{D}(w_1)| \cdot |\mathcal{D}(w_2)|} \cdot \sum_{(w_1', w_2') \in \mathcal{D}(w_1) \times \mathcal{D}(w_2)} sim(\overrightarrow{w_1'}, \overrightarrow{w_2'})$$

Similar to the average setup, prediction using centroid first averages the vectors of each word from the derivational class to obtain an average representation, whose similarity with the centroid from the second class is determined:

$$\pi_{\text{DM}_{\text{DE}}^{\text{cent}}\text{deriv}}(w_1, w_2) = sim(c(\mathcal{D}(w_1), c(\mathcal{D}(w_2))))$$

where $c(\mathcal{D}(w))$ is the centroid for the derivational class of w:

$$c(\mathcal{D}(w)) = \frac{1}{|\mathcal{D}(w)|} \cdot \sum_{w_i \in \mathcal{D}(w)} \vec{w_i}$$

The concept here is that the centroid provides a generalized meaning for the derivational class of a word similar to the prototype models described in the previous chapter.

Simple models. We consider both random and frequency baselines[9] in addition to the single DSM and SDSM wordsim models as points of comparison.

7.4.4. Results

Both experiments exhibit similar patterns across the models. First of all, the uninformed baselines – Random and Frequency – perform badlywhereas, by contrast, the single word-based DSMs already show marked increases in performance. Most encouragingly, we see that in the multilingual setting, performance metrics are comparable to the highest single-model values (correlation of .47 vs. .49 for Exp. 1, accuracy of .59 vs. .63 for Exp. 2) while at the same time boosting coverage to near-optimal levels (98% in Exp. 1 and 97% in Exp. 2). The goal of the approaches to combining the models is thus borne out: The coverage is higher than that of any single model – notably, than both the previously highest-coverage SDSMs, the DM_{DE} models, as well as the overall highest coverage of the BOW models while the performance is increased.

Experiment 1 shows a clear preference for Max over Backoff; while the coverage increases dramatically from approximately .50 and .60 for our single SDSMs to close to 1 for both multilingual modalities, the correlation values for Max remain on par with the highest performing SDSM ($\text{DM}_{\text{EN}\rightarrow\text{DE}}$) while we see a decrease for the Backoff scheme (.49 to .41).

Experiment 2 evidences a slightly less obvious distinction between the two combination options, the fact that we can combine the two single-source SDSMs to

[9] A random baseline for Exp. 1 by definition gives a correlation of 0 and is thus omitted.

Model	r	r_{cov}	cov
Baselines and single models			
Frequency	.13	.13	1
BOW	.34	.34	.97
DM_{DE}	.38	.43	.60
$DM_{EN \to DE}$	.33	.49	.49
Combined SDSM models with BOW ($DM_{EN \to DE} \to DM_{DE} \to BOW$)			
$\mathfrak{M}^{backoff}$	.40	.41	**.98**
$\mathfrak{M}^{max}$	**.49**	.50	**.98**
Other SDSM combinations			
$\mathfrak{M}^{max}_{DM_{EN \to DE} > DM_{DE}}$	.42	.47	.69
DM^{deriv}_{DE} [PSZ13]	-	.47	.89
Models from the literature			
LinGN	NA	.50	.26
JCGN	NA	.52	.33
JCGN+PLWP	NA	**.59**	.33

Table 7.3.: Exp. 1: Results for baselines and individual models (top), smoothed models (middle) and literature (bottom). Best results per column shown in bold-face.

Model	acc	acc_{cov}	cov
Baselines and single models			
Random	.25	.25	1
Frequency	.31	.31	1
BOW	.52	.53	.95
DM_{DE}	.48	.53	.84
$DM_{EN \rightarrow DE}$	.46	.61	.58
Combined SDSM models with BOW ($DM_{EN \rightarrow DE} \rightarrow DM_{DE} \rightarrow BOW$)			
$\mathfrak{M}^{backoff}$	.56	.57	**.97**
$\mathfrak{M}^{max}$	**.57**	.59	**.97**
Other SDSM combinations			
$\mathfrak{M}^{max}_{DM_{EN \rightarrow DE} > DM_{DE}}$	.55	.59	.89
DM^{deriv}_{DE} [PSZ13]	.44	.51	.87
Models from the literature			
Lin_{dist}	NA	.52	.45
HPG	NA	**.77**	.22
RPG	NA	.69	.27
JC	NA	.44	.36

Table 7.4.: Exp. 2: Results for baselines and individual models (top), combined models (middle) and literature (bottom). Best results per column shown in bold-face.

achieve near-highest accuracy (.59 for $\mathfrak{M}^{\max}$ vs. .61 for $\text{DM}_{\text{EN}\rightarrow\text{DE}}$) with almost complete coverage (.97) is encouraging due to the already high coverage of DM_{DE} of .84.

The improvement in performance attests to a level of complementarity of the information present in our mono- and cross-lingual models. While the differences among the multilingual models are small, in both experiments the maximum combination strategy edges out the backoff models, performing best overall.

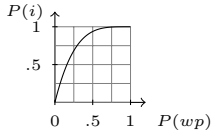
The three-way Max and Backoff models significantly outperform the baselines and single SDSMs (Exp. 2: $p < .001$).

At the same time, a two-way SDSM max-combination in both experiments shows how the positive effect on performance comes from the SDSMs while the higher coverage is gained through combining also with the simple BOW space: In Exp. 1, the two-way combined model already achieves an r_{cov} score of .47 compared to .50 for the three-way combined model, while only reaching .69 coverage, a near 30 percentage points lower than the three-way model; in Exp. 2, the two-way model already reaches the same accuracy as the three-way model, .59 at a commendable level of coverage: .89. One reason for the higher coverage in general in Exp. 2 vs. Exp. 1 is that for an item to be out of coverage in Exp. 2, the model must be out of coverage comparing the target word with all four synonym candidates. Re-interpreting the coverage as probability of a given word pair being covered $P(wp)$, we can give a rough estimate of the probability of an item being covered $P(i)$ as a function of $P(wp)$:

Analytically:

$$\begin{aligned} cov \stackrel{\text{def}}{=} P(\text{item } i \text{ covered}) &= 1 - \prod_{k=1}^{4}(1 - P(\text{word pair } k \text{ covered})) \\ &\approx 1 - (1 - P(wp))^4 \end{aligned}$$

Graphically:



This a word-pair coverage of 50% would already lead to a per-item coverage of over 93%. Obviously, this is overly simplified: In reality, there are correlations in coverage within an item (as the target word remains constant over k for each i) which would reduce item coverage.

7.4.5. Qualitative analysis of results on Exp. 1 and 2.

The largest improvement in Exp. 1 is in coverage, so we will now take a look at the items in Gur350 that are covered by the combined models which previously were not.

First, we identify a type of word pairs that are unlikely to be covered by a translation lexicon which can include proper nouns, such as *Berlin – Berlin-Kreuzberg*, *Benedetto – Benedikt*. Short of utilizing a method for determining their similarity as strings, our models would require a vast amount of world knowledge to generate predictions for such items as: *Ratzinger – Papst* (pope) which – in our view – has a better chance of occurring in a target language corpus. Indeed, we find that this pair is also not covered by our cross-lingual models but can be assigned a relatedness score by the monolingual (S)DSMs: DM_{DE}, and BOW assign it the similarities .23, and .89, respectively.

Next we compare the quality of predictions between the three-way max and backoff models with respect to the human judgments in the dataset. To achieve this, we first define the **prediction difference** for a model x on item i:

$$\Delta x(i) = \pi_x(i) - \text{judg}_i$$

that is the difference between the prediction of that model to the human judgment for that item. Smaller prediction differences correspond to more accurate predictions.

Using this metric, we can visualize the performance difference between two models in two dimensions. Each item will correspond to a point with the two models' predictions making up the x- and y-coordinates, respectively. Finally, we connect this point with the first diagonal $y = x$ at the point corresponding to the human judgment for that item. Then the slope $\frac{\Delta y}{\Delta x}$ (cf. Figure 7.5) of that segment shows us whether model x or y had more accurate prediction.

In Figure 7.6 we see that, in general, the max-combined model has smaller-valued prediction differences than backoff.[10] Examples for which the Δ-values differ maximally are shown in Figure 7.6, these include pairs for which $\Delta\text{backoff} < \Delta\text{max}$ such as *Projekt – Aktion* (project – action/activity), *Formulierung – Stiftung*

[10]Note by contrast, that by definition max will produce estimates equal to or greater than the backoff model.

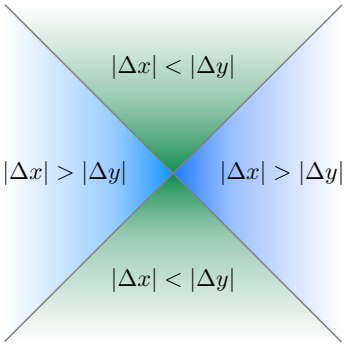


Figure 7.5.: Slopes show relative prediction differences for two models x and y. Segments with $\frac{\Delta y}{\Delta x} < 1$ mean that model y's prediction is closer than model x's. Thus, blue areas correspond to better performance of model y and green areas show where model x has more accurate predictions.

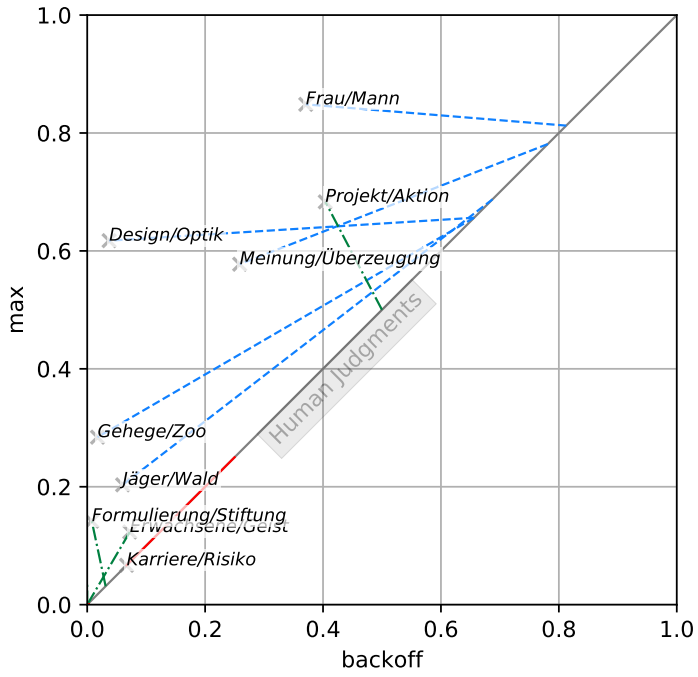


Figure 7.6.: Examples of prediction differences for backoff and max models in two-way combination. Each color of the line segments denotes the relative quality in predictions (Green: backoff better than max, Blue: max better than backoff, Red: identical).

(formulation – foundation), *Erwachsene* – *Geist* (adult – spirit/ghost); those for which Δmax $<$ Δbackoff *Frau* – *Mann* (woman – man), *Design* – *Optik* (design – optics), *Meinung* – *Überzeugung* (opinion – conviction), *Gehege* – *Zoo* (corral – zoo), *Jäger* – *Wald* (hunter – forrest); and an example where Δmax $=$ Δbackoff $<$ judg: *Karriere* – *Risiko* (career – risk).

Analysis of example. By way of example, we investigate the difference in relatedness predictions of the pair *Frau* – *Mann* (woman – man). The estimated relatedness is lower – and farther from human judgment – using backoff (i.e. cross-lingual) as compared to max combination. We compare the top contexts for the two words in DM_{EN} (the source for $\text{DM}_{\text{EN}\rightarrow\text{DE}}$) as well as DM_{DE}. We determine the top link types for $w_1 \in \{man, woman\}$ as those with the highest *LMI* scores (i.e. $arg\,max_l\,[max_{w_2}\sigma(\langle w_1\ l\ w_2\rangle)]$), and then within each listed link type, the top-scored w_2 are listed in Tables 7.5 and 7.6.

Among their top contexts, in DM_{EN}, *man* and *woman* have more distinct than shared typical contexts (i.e. as subjects, objects and so on); whereas in the case of DM_{DE}, the top subject and object contexts for *Mann* and *Frau* are more similar than distinct, i.e. only on a few of their top contexts do they differ. Table 7.5 shows how the top links share more top context words, with lower ranking links showing the opposite tendency. The reverse is the case in Table 7.6 where the top links evidence more unique context words and only among the later links do we see the relative number of shared context words take over.[11]

[11] NB: Recall that the weights used to determine 'typical' links and context words here are the local mutual information scores (cf. Chapter 3). This means that precedence in this list is due to a confluence of absolute frequency and informativity.

SUBJ_TR	*Mann*$_n$	*sitzen*$_v$ (sit) *gehen*$_v$ (go/walk) *halten*$_v$ (hold)
	Frau$_n$	*bekommen*$_v$ (receive) *erhalten*$_v$ (receive) *fühlen*$_v$ (feel)
	both	*tragen*$_v$ (carry) *stehen*$_v$ (stand) *machen*$_v$ (make) *kommen*$_v$ (come) *sagen*$_v$ (say) *nehmen*$_v$ (take) *sehen*$_v$ (see)
SUBJ_INTR	*Mann*$_n$	*treten*$_v$ (step) *heißen*$_v$ (be called)
	Frau$_n$	*leiden*$_v$ (suffer) *erkranken*$_v$ (fall ill)
	both	*gehen*$_v$ (go/walk) *sterben*$_v$ (die) *stehen*$_v$ (stand) *sitzen*$_v$ (sit) *kommen*$_v$ (come) *arbeiten*$_v$ (work) *sagen*$_v$ (say) *leben*$_v$ (live)
OBJ	*Mann*$_n$	*erschießen*$_v$ (shoot to death) *finden*$_v$ (find)
	Frau$_n$	*vergewaltigen*$_v$ (rape) *bringen*$_v$ (bring)
	both	*geben*$_v$ (give) *suchen*$_v$ (search) *lieben*$_v$ (love) *heiraten*$_v$ (marry) *kennen*$_v$ (know) *töten*$_v$ (kill) *sehen*$_v$ (see) *treffen*$_v$ (meet)
IOBJ	*Mann*$_n$	*vorbehalten*$_v$ (reserve) *erzählen*$_v$ (tell) *fallen*$_v$ (fall) *folgen*$_v$ (follow)
	Frau$_n$	*machen*$_v$ (make) *danken*$_v$ (thank) *gehen*$_v$ (go/walk) *ermöglichen*$_v$ (enable)
	both	*geben*$_v$ (give) *stehen*$_v$ (stand) *begegnen*$_v$ (encounter) *gelingen*$_v$ (succeed) *helfen*$_v$ (help) *sagen*$_v$ (say)
VERB	*Mann*$_n$	*Frau*$_n$ (woman) *Mädchen*$_n$ (girl) *Kopf*$_n$ (head) *Arbeit*$_n$ (work) *Geld*$_n$ (money) *Wort*$_n$ (word) *Hand*$_n$ (hand)
	Frau$_n$	*Hilfe*$_n$ (help) *Rolle*$_n$ (role) *Mann*$_n$ (man) *Unterstützung*$_n$ (support) *Wahlrecht*$_n$ (suffrage) *Kopftuch*$_n$ (headscarf) *Recht*$_n$ (right)
	both	*Kind*$_n$ (child) *Leben*$_n$ (life) *Weg*$_n$ (way)
VERB^{-1}	*Mann*$_n$	*Gott*$_n$ (God) *Polizeibeamter*$_n$ (police officer) *Heinrich*$_n$ (Heinrich) *Zeuge*$_n$ (witness) *Thomas*$_n$ (Thomas)
	Frau$_n$	*Ehemann*$_n$ (husband) *Adam*$_n$ (Adam) *Vater*$_n$ (father) *Stadt*$_n$ (city) *Paul*$_n$ (Paul)
	both	*Frau*$_n$ (woman) *Mann*$_n$ (man) *Beamter*$_n$ (official) *Polizist*$_n$ (policeman) *Polizei*$_n$ (police)
IN	*Mann*$_n$	*Gefängnis*$_n$ (prison) *Rollstuhl*$_n$ (wheelchair) *Wohnung*$_n$ (apartment) *Anzug*$_n$ (suit) *Uniform*$_n$ (uniform)
	Frau$_n$	*Gesellschaft*$_n$ (society) *Führungsposition*$_n$ (leadership position) *Bereich*$_n$ (area) *Land*$_n$ (land/country) *Beruf*$_n$ (profession)
	both	*Haus*$_n$ (house) *Jahr*$_n$ (year) *Bett*$_n$ (bed) *Leben*$_n$ (life) *Alter*$_n$ (age)
MIT	*Mann*$_n$	*Frau*$_n$ (woman) *Hut*$_n$ (hat) *Gesicht*$_n$ (face) *Bart*$_n$ (beard) *Hammer*$_n$ (hammer) *Brille*$_n$ (glasses) *Name*$_n$ (name)
	Frau$_n$	*Kind*$_n$ (child) *Behinderung*$_n$ (handicap) *Kinderwunsch*$_n$ (wish to have children) *Tochter*$_n$ (daughter) *Brustkrebs*$_n$ (breast cancer) *Mann*$_n$ (man) *Kopftuch*$_n$ (headscarf)
	both	*Stimme*$_n$ (voice) *Haar*$_n$ (hair) *Auge*$_n$ (eye)

Table 7.5.: Most highly associated contexts (links and words) for *Mann*$_n$ and *Frau*$_n$ in DM$_{DE}$. For each link, we first list those context words from the top 10 highly associated exclusive to either *Mann*$_n$ or *Frau*$_n$, then their shared contexts.

Qualitative analysis of Exp. 2. Inspecting Exp. 2 for synonyms that were correctly detected by $\mathfrak{M}^{\max}$ but not $\mathrm{DM}_{\mathrm{DE}}$ we find a number of words of foreign origin:

Nouns	*Couscous* (couscous), *Albino* (albino)
Adjectives	*kursorisch* (cursory), *süffisant* (smug)
Verbs	*erodieren* (erode), *moussieren* (fizz)

These words tend to be rare in the German corpus with number of occurrences ranging from 528 for *erodieren* to 6 for *moussieren*. Such words in the form of technical terms, slang, elevated register or loan words which will more commonly tested in vocabulary building exercises such as those upon which the RDWP dataset is built. They also exhibit a low level of ambiguity as measured by number of translations in dict.cc – in these examples, translational variance values are either 1 or 2. Finally, we find that their English translations are often more frequent, which gives the cross-lingual model an advantage and improves the quality of their representations.

Summary. There is a difference in the relative performances in the two experiments.

On word relatedness, prediction combination has a larger impact, and max shows a clearer improvement over the single models than the backoff combination. Considering the task in Exp. 1 which is correlation of human relatedness judgments with model relatedness predictions in which each single prediction can affect the entire outcome, $r = \frac{E[X-\mu_X]\cdot E[Y-\mu_Y]}{\sigma_X \cdot \sigma_Y}$. In Exp. 2, the contribution of an item in the classification problem to the final score is an aggregate measure on the correct ordering of one word-pair comparison among four. The margin of the decision boundaries themselves are irrelevant. We should thus expect there to be less sensitivity to fluctuations in prediction quality – which is what we observe in Exp. 2: On Gur350 backoff- and max-combined predictions differ on over close to half of the items (155 of 350 pairs), while the predicted synonyms differ only for in 5% of cases (52 of 984 items).

Figure 7.7 graphically shows the effect of incrementally adding the higher coverage models into the combination process starting from $\mathrm{DM}_{\mathrm{EN}\to\mathrm{DE}}$ then adding $\mathrm{DM}_{\mathrm{DE}}$ and finally BOW. The plots illustrate how the quality of the models can be

SBJ_INTR	man_n	$walk_v$ $stand_v$ $think_v$ $live_v$ $look_v$
	$woman_n$	$wear_v$ $become_v$ $work_v$ $want_v$ $feel_v$
	both	do_v say_v go_v die_v $come_v$
OBJ	man_n	$send_v$ $arrest_v$ $beat_v$ $know_v$ $call_v$
	$woman_n$	$rape_v$ $affect_v$ $treat_v$ $help_v$ $employ_v$
	both	$marry_v$ $meet_v$ $kill_v$ see_v $find_v$
SBJ_TR	man_n	$leave_v$ see_v $lose_v$ $hold_v$
	$woman_n$	use_v $play_v$ $receive_v$ $experience_v$
	both	do_v $wear_v$ get_v $give_v$ $take_v$ $make_v$
VERB	man_n	$wife_n$ $girl_n$ car_n $house_n$ $thing_n$ job_n $life_n$
	$woman_n$	$child_n$ $pregnancy_n$ $medal_n$ $dress_n$ $husband_n$ $veil_n$ $abortion_n$
	both	$injury_n$ $woman_n$ man_n
NMOD	man_n	$coach_n$ sin_n $wage_n$ $chorus_n$ man_n $scull_n$ $shoe_n$
	$woman_n$	$organization_n$ $activist_n$ $team_n$ $championship_n$ $league_n$ $college_n$ $dress_n$
	both	$tournament_n$ $magazine_n$ $champion_n$
$NMOD^{-1}$	man_n	$brave_j$ $honest_j$ $rich_j$ gay_j $wealthy_j$ $tall_j$ $wise_j$ $good_j$
	$woman_n$	$married_j$ $sexy_j$ few_j $beautiful_j$ $poor_j$ $pregnant_j$ $many_j$ $attractive_j$
	both	old_j $young_j$
FOR	man_n	$pray_v$ $make_v$ $search_v$ $wait_v$ $feel_v$
	$woman_n$	$reserve_v$ $provide_v$ $work_v$ $design_v$ $fall_v$
	both	do_v $care_v$ $leave_v$ $gift_n$ $look_v$
OF	man_n	$biography_n$ $memory_n$ $tale_n$ $speak_v$
	$woman_n$	$rape_n$ $proportion_n$ $murder_n$ $stereotype_n$
	both	$picture_n$ $portrait_n$ $consist_v$ $think_v$ $image_n$ $portrayal_n$
WITH	man_n	go_v $meet_v$
	$woman_n$	$relationship_n$ $involve_v$
	both	$associate_v$ $sleep_v$ $flirt_v$ $share_v$ $work_v$ $live_v$ $fall_v$ $deal_v$
AS	man_n	$regard_v$ act_v
	$woman_n$	$dress_v$ $depict_v$
	both	$appear_v$ $portray_v$ $describe_v$ $live_v$ see_v $refer_v$ $know_v$ $disguise_v$
IOBJ	man_n	$warn_v$ $cost_v$ $order_v$
	$woman_n$	$advise_v$ $deny_v$ $grant_v$
	both	$teach_v$ $offer_v$ ask_v $give_v$ $tell_v$ $persuade_v$ $urge_v$

Table 7.6.: Most highly associated contexts (links and words) for man_n and $woman_n$ in DM_{EN}. For each link, we first list those context words from the top 10 highly associated exclusive to either man_n or $woman_n$, then their shared contexts.

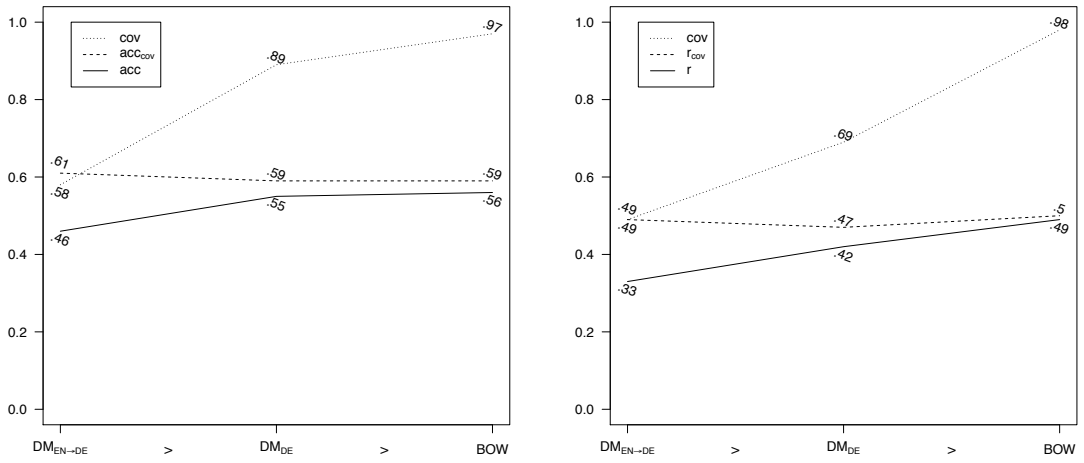


Figure 7.7.: Performance of incremental combination ($DM_{EN \to DE}$, $DM_{EN \to DE}$ →DM, $DM_{EN \to DE}$ →DM→BOW) using score maximization for the word relatedness (left) and synonym choice (right) tasks

maintained or improved while dramatically improving coverage. The improvements to coverage from DM_{DE} and BOW in Exp. 1 are comparable, while for Exp. 2, the major increase comes from adding the monolingually induced DM_{DE}, it thus suffices to have some monolingual data to be estimate similarities in a setting such as the word power vocabulary building exercise. This shows the robustness of the max combination setup: The resulting multilingual models successfully combine the high coverage of monolingual models with the high quality of cross-lingual models.

7.5. Evaluation of Summarization on Croatian Data

In addition to building and testing our models on German which is in the same subgroup of languages as English, i.e., the Germanic languages which also include Dutch and the Scandinavian languages among others, we test our methods on a less related language, Croatian. Similar to the analyses outlined in the preceding sections, we combine the monolingually induced SDSM DM_{HR} (Šnajder et al., 2013) with a cross-lingual $DM_{EN \to HR}$.

Evaluations are carried out on word relatedness and synonym choice experiments as previously done for German.

BOW model. The simple DSM acting as a baseline in these experiments is based on a 5-word context window and has been reduced to 500 dimensions using PCA.

Freq baseline. The frequency baselines assign either the maximum frequency for a word pair in the word relatedness task, or choose the candidate with the highest frequency as synonym in the synonym choice task.

$\mathfrak{M}^{\text{backoff}}$ and $\mathfrak{M}^{\text{max}}$. We combine according to the methodology described above (cf. Section 7.3) the Croatian single-language source SDSMs as well as the simple DSM BOW model to obtain our multilingual models and compare them.

7.5.1. Word relatedness (CroSemRel)

Dataset. The word relatedness dataset for Croatian[12] (Janković, Šnajder, and Bašić, 2011) covers 450 word pairs annotated by 6 annotators (a subset of 12 annotators with highest agreement).

Results. Table 7.7 lists the Pearson correlation values on the word relatedness task.

The cross-lingual DM has markedly higher correlation compared to DM_{HR} (.64 vs. .45) and only has a slightly lower coverage (.96 vs. 1). The top performing models are $\text{DM}_{\text{EN}\rightarrow\text{HR}}$, $\mathfrak{M}^{\text{backoff}}$ and $\mathfrak{M}^{\text{max}}$ ($r = .64$ and .63). Interestingly, the fully backoff-combined model actually shows a nominally better correlation with human judgments than the fully max-combined one. All other cross- and multilingual models perform similarly – with the exception of backing off from DM_{HR} first results in an r value of .45. We can thus see that inaccurate estimates for BOW have 'overpowered' the generally higher quality ones. Figure 7.8 shows differences in

[12]Obtained from: http://takelab.fer.hr/data/crosemrel450/

Model	r	r_{cov}	cov
Freq	0.16	0.16	1.00
BOW	0.26	0.26	1.00
DM_{HR}	0	0	0
$\text{DM}_{\text{EN}\rightarrow\text{HR}}$	0	0	0
Fully combined SDSM models			
$\mathfrak{M}^{\text{backoff}}_{\text{DM}_{\text{EN}\rightarrow\text{HR}}\rightarrow\text{DM}_{\text{HR}}\rightarrow\text{BOW}}$	0.63	0.63	1.00
$\mathfrak{M}^{\text{max}}_{\text{DM}_{\text{EN}\rightarrow\text{HR}}>\text{DM}_{\text{HR}}>\text{BOW}}$	0.61	0.61	1.00
Other combined models			
$\mathfrak{M}^{\text{backoff}}_{\text{DM}_{\text{EN}\rightarrow\text{HR}}\rightarrow\text{DM}_{\text{HR}}}$	0.63	0.63	1.00
$\mathfrak{M}^{\text{backoff}}_{\text{DM}_{\text{HR}}\rightarrow\text{DM}_{\text{EN}\rightarrow\text{HR}}}$	0.45	0.45	1.00
$\mathfrak{M}^{\text{max}}_{\text{DM}_{\text{EN}\rightarrow\text{HR}}>\text{DM}_{\text{HR}}}$	**0.64**	**0.64**	1.00

Table 7.7.: Performance on Croatian word relatedness task.

estimates for $\mathfrak{M}^{\text{max}}$ with and without BOW on the 105 word pairs (23%) on which their relatedness predictions differ.

We can identify four word pairs that are outliers whose similarities are significantly overestimated by $\mathfrak{M}^{\text{max}}_{\text{w/ BOW}}$: *djeca/državni* (children/national), *trenutak/interes* (moment/interest), *američki/srijeda* (American/Wednesday), *obzir/dodati* (consideration/add). Note that all of these pairs except *trenutak/interes* cross parts of speech. As a consequence, we would expect lower estimates from a structured model that an unstructured BOW, which is what we find. When we remove these outliers we obtain to a correlation for $\mathfrak{M}^{\text{max}}_{\text{DM}_{\text{EN}\rightarrow\text{HR}}>\text{DM}_{\text{HR}}>BOW}$ of .64, putting it on par with the top performing model.[13]

In addition, we also observe mild frequency effects. First, the total frequency of these outliers is below average (gray / grayish blue color). Second, the proportional difference in frequency is relatively small (smaller points); this could mean we get spuriously high similarities with the inclusion of BOW due to chance

[13]This sensitivity to outliers can be reduced by using Spearman's ρ rank correlation over Pearson's r values. In terms of performance measured in ρ, we indeed find $\mathfrak{M}^{\text{max}}$ performs equally well with or without BOW ($\rho = .55$ vs. .56) while outperforming $\mathfrak{M}^{\text{backoff}}$ in either condition ($\rho = .50$). However, we report r to maintain consistency with previous chapters as well as prior and related work.

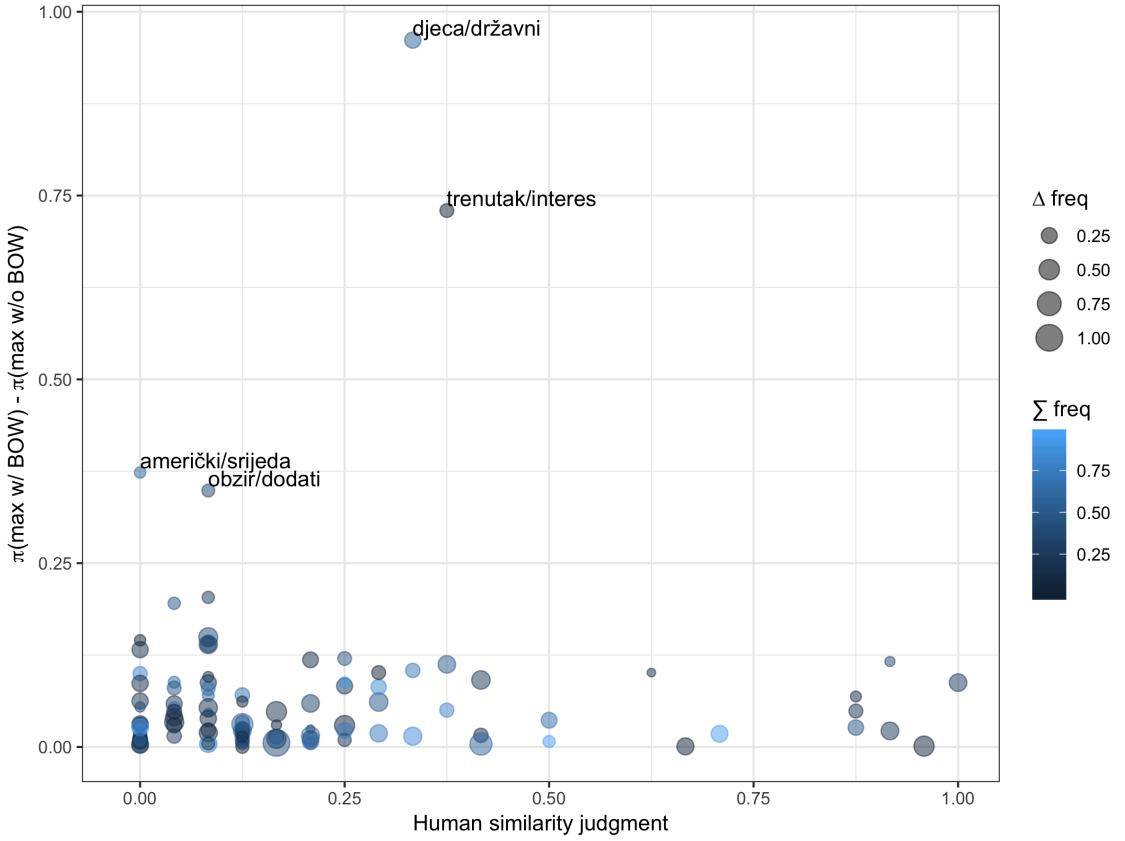


Figure 7.8.: Differences in estimates of $\mathfrak{M}^{\max}_{\text{DM}_{\text{EN}\to\text{HR}}>\text{DM}_{\text{HR}}>BOW}$ ('w/ BOW') and $\mathfrak{M}^{\max}_{\text{DM}_{\text{EN}\to\text{HR}}>\text{DM}_{\text{HR}}}$ ('w/o BOW'): The x values correspond to human relatedness judgments for each word pair, the y values are the differences between relatedness predictions for the respective models. Each point's area corresponds to the normalized difference in corpus frequencies for the pair of words $\left(\Delta \text{ freq} = \frac{|\text{freq}_1 - \text{freq}_2|}{\text{freq}_1 + \text{freq}_2}\right)$ while its color shows the total frequency $(\text{freq}_1 + \text{freq}_2)$ of a word pair relative to the maximum total frequency (light blue means the word pair has close to maximal total frequency on the dataset: $max(\text{freq}_1 + \text{freq}_2) = 2.2 \cdot 10^6$). Only word pairs for which the estimates differ are considered, i.e. $\pi(\mathfrak{M}^{\max}_{\text{w/ BOW}}) \neq \pi(\mathfrak{M}^{\max}_{\text{w/o BOW}})$.

frequency profile similarity. Finally, the lighter and larger points are located where $\pi(\text{w/ BOW}) \approx \pi(\text{w/o BOW})$. Additional experiments indeed show considering only pairs above a certain threshold θ of proportional frequency difference and total frequency lead to higher correlation values. The maximum correlation values attainable for the full max multilingual model are $r_{cov} = .83$ when thresholding

using total frequency ($\theta = .53, n = 13, cov = .03$), and $r_{cov} = .75$ when thresholding using proportional frequency differences ($\theta = .89, n = 30, cov = .07$).

7.5.2. Synonym Choice (CroSyn)

Dataset. The CroSyn dataset[14] (Karan et al., 2012; Šnajder et al., 2013) was compiled to conform to the structure of the RDWP dataset: One target word is paired with four candidate words, of which one is the true synonym and three are distractors (cf. Section 3.4.2). The information it is based upon is the machine-readable Croatian dictionary compiled by Anić (2003). On the basis of the synonym links present in the dictionary as well as a corpus to ensure that words included have attested use, 1000 items for each part of speech (noun, verb and adjective) were extracted. Synonyms with a high degree of character overlap were excluded (these were assumed to be morphologically related), and the distractor items were selected from the corpus at random with the restriction that they were not linked to either target or synonym via transitive synonymy links of any length.

Results. Table 7.8 shows the performance of our mono-, cross- and multilingual models on the Croatian synonym choice task.

As in the previous task, coverage values are all close to or equal to 100%. The frequency baseline Freq performs worse than random ($.22 < .25$). BOW already reaches comparable accuracy to the monolingual DM_{HR} ($.61$ and $.66$, respectively). Both have basically 100% coverage (slight difference between 99.6% & 99.7% coverage, respectively). The final single model $\text{DM}_{\text{EN}\rightarrow\text{HR}}$ has a higher accuracy on covered items but due to its coverage of only 71% it has a lower overall accuracy than BOW or DM_{HR} ($.49$ versus $.61$ and $.66$ respectively).

Due to the already high coverage values of the single SDSM models, one round of combination of $\text{DM}_{\text{EN}\rightarrow\text{HR}}$ with DM_{HR} is sufficient to reach full coverage with max proving to be the superior combination modality over backoff.

The following contingency table contains the frequencies of word pairs for which we have differential predictions across max/backoff.

[14] Obtained from: http://takelab.fer.hr/data/crosyn/

Model	acc	acc_{cov}	cov
Freq	0.22	0.22	1.00
BOW	0.61	0.62	1.00
DM_{HR}	0.66	0.66	1.00
$\text{DM}_{\text{EN}\rightarrow\text{HR}}$	0.49	0.68	0.71
Fully combined SDSM models			
$\mathfrak{M}^{\text{backoff}}_{\text{DM}_{\text{EN}\rightarrow\text{HR}}\rightarrow\text{DM}_{\text{HR}}\rightarrow\text{BOW}}$	0.65	0.65	1.00
$\mathfrak{M}^{\text{max}}_{\text{DM}_{\text{EN}\rightarrow\text{HR}}>\text{DM}_{\text{HR}}>\text{BOW}}$	**0.71**	**0.71**	1.00
Other combined models			
$\mathfrak{M}^{\text{backoff}}_{\text{DM}_{\text{EN}\rightarrow\text{HR}}\rightarrow\text{DM}_{\text{HR}}}$	0.65	0.65	1.00
$\mathfrak{M}^{\text{backoff}}_{\text{DM}_{\text{HR}}\rightarrow\text{DM}_{\text{EN}\rightarrow\text{HR}}}$	0.66	0.66	1.00
$\mathfrak{M}^{\text{max}}_{\text{DM}_{\text{EN}\rightarrow\text{HR}}>\text{DM}_{\text{HR}}}$	**0.71**	**0.71**	1.00

Table 7.8.: Performance on Croatian synonym choice task

	n	j	v
$\sigma(\mathfrak{M}^{\text{max}}) > \sigma(\mathfrak{M}^{\text{backoff}})$	108	123	187
$\sigma(\mathfrak{M}^{\text{max}}) < \sigma(\mathfrak{M}^{\text{backoff}})$	83	83	86

A χ^2-test showed that part of speech is associated with max outperforming backoff ($\chi^2 = 7.7531$, $df = 2$, $p < 0.05$), which means that depending on which part of speech a word pair is taken from, a max- or backoff-based combination strategy can be expected to perform better or worse.

Figure 7.9 investigates whether there is a frequency effect that might underlie this association and we investigate the difference between the fully combined $\mathfrak{M}^{\text{max}}$ and $\mathfrak{M}^{\text{backoff}}$ models using density plots of the target words' log-transformed frequencies in fHRWAC. One-tailed t-tests were performed to check for differences of means.

Subfigure (a) shows that $\mathfrak{M}^{\text{max}}$ predicts better on higher frequency targets and indeed we find a significant difference ($p < .01$) in mean log-transformed target frequencies for the two groups:

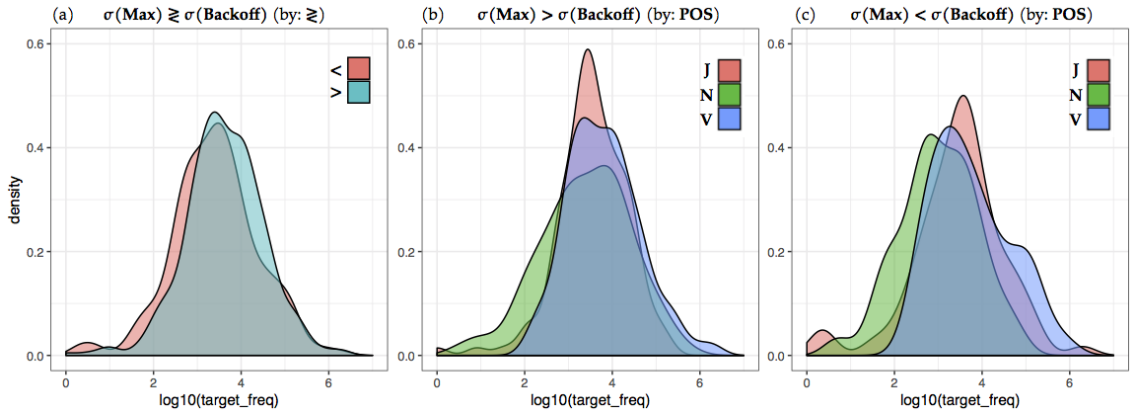


Figure 7.9.: Density plots of target frequencies for different groupings of model scores for $\mathfrak{M}^{\max}_{\text{DM}_{\text{EN}\to\text{HR}}>\text{DM}_{\text{HR}}>BOW}$ ('Max') and $\mathfrak{M}^{\text{backoff}}_{\text{DM}_{\text{EN}\to\text{HR}}\to\text{DM}_{\text{HR}}\to BOW}$ ('Backoff').

$$\mu|_{M>B}(\text{LTF}) > \mu|_{M<B}(\text{LTF})$$

where LTF = $\log$(target freq) and $M \gtrless B = \sigma(\mathfrak{M}^{\max}) \gtrless \sigma(\mathfrak{M}^{\text{backoff}})$. Subfigure (b) shows the significant ($p < .001$) differences in target frequencies for which max outperforms backoff:

$$\mu|_{\substack{M>B \\ POS=v}}(\text{LTF}) > \mu|_{\substack{M>B \\ POS=j}}(\text{LTF})$$
$$\mu|_{\substack{M>B \\ POS=v}}(\text{LTF}) > \mu|_{\substack{M>B \\ POS=n}}(\text{LTF})$$

and Subfigure (c) reveals that on items for which backoff performs better ($p < .05$):

$$\mu|_{\substack{M<B \\ POS=j}}(\text{LTF}) > \mu|_{\substack{M<B \\ POS=n}}(\text{LTF})$$

We should thus be able to artificially boost the accuracy on covered items with a frequency threshold. In Table 7.9 we list the maximal acc_{cov} for thresholded target

frequencies.

model	θ	acc	acc_{cov}	cov
$\text{DM}_{\text{EN}\rightarrow\text{HR}}$	$6 \cdot 10^4$	.09	.71	.12
$\mathfrak{M}^{\text{backoff}}$	$6 \cdot 10^4$	.08	.69	.12
$\mathfrak{M}^{\max}$	$1.4 \cdot 10^5$	.04	.74	.06

Table 7.9.: Frequency thresholding on Croatian synonym choice task.

While the other models' accuracy values shifted dramatically with thresholding, interestingly, $\text{DM}_{\text{EN}\rightarrow\text{HR}}$ remained very stable at .68 across frequency thresholds. We have shown that $\text{DM}_{\text{EN}\rightarrow\text{HR}}$ and the combined models exhibit differential performance across parts of speech which is in line with the analysis of Šnajder et al. (2013) for DM_{HR}. We have gone further, however, and shown there exists a connection with the frequency of targets: Even though the frequency model itself does not contain enough information to provide accurate predictions, it is possible to use frequency information to fine-tune the performance of combined models.

In the following section, we outline an argument for why the maximum combination method should be expected to perform more reliably in the setting of combining distributional models.

7.6. Underestimation Hypothesis (UEH)

In this section, we will argue and give evidence for the intuition that the noise inherent in the construction process of a model is unlikely to increase vector similarity by chance. Thus, in our case, a semantic space can be expected to be biased to underestimate rather than overestimate relatedness in a distributional space; this is what we term the **underestimation hypothesis**. We first develop this intuition further via geometrical reasoning and then confirm the conclusions drawn with an empirical study.

Geometrical argument. Distributional models, as approximations of the true underlying semantic information – essentially contain errors in the true co-occurrence frequencies or association magnitudes. If the for a given word w is the ideal vector

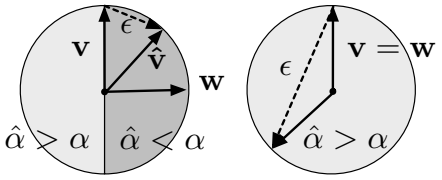


Figure 7.10.: Underestimation hypothesis: ideal and empirical vectors $(\vec{v}, \vec{\hat{v}})$, point of comparison $(\vec{w})$, noise vector $(\vec{\epsilon})$. Segments of the hypersphere where angle decreases (dark grey) and increases (light grey). Left: $\alpha = 90°$ (lower $\hat{\alpha}$), Right: $\alpha = 0°$ (higher $\hat{\alpha}$).

$\vec{v}$, then the actual empirical vector we find in the space can be written as $\vec{\hat{v}} = \vec{v} + \vec{\epsilon}$ for some noise vector $\vec{\epsilon}$. Now we can investigate the effect of this noise on the cosine similarity – if UEH holds, then a decrease would be the more likely scenario, i.e. $cos(\vec{v}, \vec{w}) > cos(\vec{\hat{v}}, \vec{w})$, with $\vec{w}$ some other semantic vector in the same space, or, in other words:

$$P(sim(\vec{v}, \vec{w}) > sim(\vec{\hat{v}}, \vec{w})) > P(sim(\vec{v}, \vec{w}) < sim(\vec{\hat{v}}, \vec{w}))$$

Take the vectors $\vec{v}$, $\vec{\hat{v}}$, and $\vec{w}$ to be normalized without loss of generality. Then each vector corresponds to a point on the unit hypersphere. The cosine similarity then decreases if and only if the angle between $\hat{\alpha}$ between $\vec{\hat{v}}$ and $\vec{w}$ – the 'empirical' angle – is larger than the 'ideal' angle α between $\vec{v}$ and $\vec{w}$. As Figure 7.10 shows, this is the case outside a hypercone spanning the angle 2α around $\vec{w}$. In the case of $\alpha = 90°$ – which is the case when the information in $\vec{v}$ and in $\vec{w}$ is disjoint – this cone is maximally wide and it is equally likely for the cosine to decrease as to increase, i.e. without making any assumptions on the distribution of the noise vector $\vec{\epsilon}$. However, for all smaller angles α, this cone shrinks meaning the likelihood of a smaller cosine increases. Then finally, in the extreme case of maximal ideal similarity (i.e. $\vec{v} = \vec{w} \Leftrightarrow \alpha = 0°$), any noise will necessarily decrease the similarity.

Experimental support for UEH. To further substantiate the validity of UEH without having access to actual ideal vectors, we offer the following experimental evidence. Inasmuch as distributional representations can be conceived of as

capturing the meaning content and variation of words in their use, it appears uncontroversial to claim that larger corpora will lead to closer-to-ideal empirical vectors. We thus propose to simulate the relationship between ideal and empirical vectors by downsampling our available corpus, defining ideal vectors as those obtained from the full corpus and empirical ones being obtained from a subset of the corpus.[15] To do this, we randomly partition the set of sentences of the SDEWAC corpus into two halves generating two 'half subspaces'. The word similarities obtained from these two subspaces are termed 'half similarities' $sim_{\frac{i}{2}}(\vec{v},\vec{w})$ ($i \in \{1,2\}$) – which will represent our empirical similarities – and those from the full space are termed 'full similarities' $sim_{\frac{1}{1}}(\vec{v},\vec{w})$. If UEH holds, it will more often be the case that the half similarities will be lower than the corresponding full sim:

$$|\{\langle\vec{v},\vec{w}\rangle \mid sim_{\frac{i}{2}}(\vec{v},\vec{w}) < sim_{\frac{1}{1}}(\vec{v},\vec{w})\}| > |\{\langle\vec{v},\vec{w}\rangle \mid sim_{\frac{i}{2}}(\vec{v},\vec{w}) > sim_{\frac{1}{1}}(\vec{v},\vec{w})\}|$$

that is, the number of pairs $\langle\vec{v},\vec{w}\rangle$ for which the half similarities are less than the full similarities will be greater.

We stratify the word pairs by frequency of the word pairs involved. Recall that on our view, vectors with more evidence are closer to the ideal vector and vice versa. We thus expect lower frequency word pairs (measured in terms of the minimum of the single words' frequencies) to evidence more underestimation, i.e. $sim_{\frac{i}{2}}(\vec{v},\vec{w}) < sim_{\frac{1}{1}}(\vec{v},\vec{w})$, or $sim_{\frac{i}{2}}(\vec{v},\vec{w}) - sim_{\frac{1}{1}}(\vec{v},\vec{w}) < 0$.

Figure 7.11 shows that the lowest frequency word pairs indeed suffer more from underestimation.

In addition to the above, we also performed an in-vivo test. Using the Gur350 word pairs (cf. Task 1 in Chapter 6) we perform a one-tailed t-test on the similarity estimates from the full via doubling of the full similarities and two half spaces and obtain a highly significant underestimation ($t = -4.3647$, $p < .0001$). A side result of note in these experiments was that the full similarities showed higher correlation with either half similarities than the half similarities did among themselves.

[15] This is similar the resampling methods in statistics, e.g. bootstrap (Efron and Tibshirani, 1993) or jackknife (Efron, 1982), in which the large sample is taken to be the true underlying distribution and subsampling is used to test hypotheses or generate confidence intervals of estimates of the larger sample.

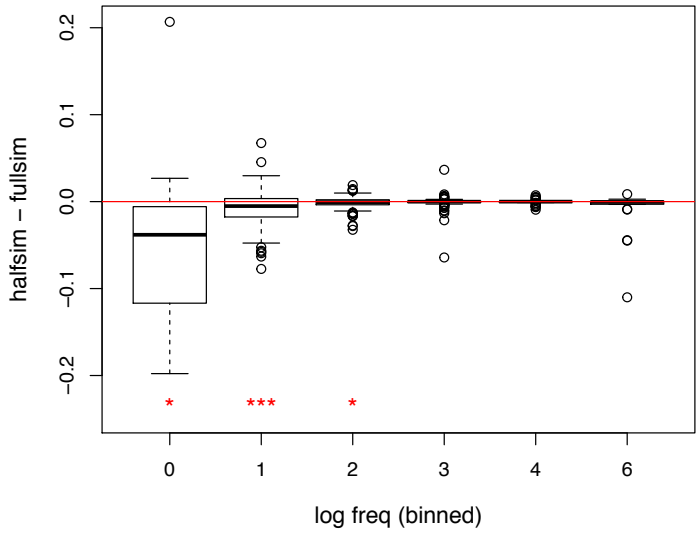


Figure 7.11.: Differences between full and half similarities by log frequency of word pairs. (Significance levels are shown for paired one-tailed t-tests within each bin: $***: p < 0.001, **: p < 0.01, *: p < 0.05$ of test of mean difference from 0.)

As a consequence of UEH, if any of the models predicts a higher relatedness, this fact can itself be viewed as evidence of a more reliable approximation of the ideal similarity and should be used to the exclusion of additional, smaller estimates. We take this as prima facie evidence for a better performance of a maximum-based combination of estimates in a distributional model.

Part IV.

Conclusion

Chapter 8.

Conclusion

Contributions. In this thesis I have investigated methods for constructing structured distributional semantic models for languages with low resource availability. My particular focus was whether viable general-purpose SDSMs such as Distributional Memories (DMs) had previously shown to be for English could be obtained for languages with either smaller text data coverage or even completely without parsed text corpora. After constructing a novel German SDSM adapting the original work on English DM by Baroni and Lenci, I developed methodologies for drastically reducing the necessary resource requirements.

The two resources for German constructed and were evaluated on numerous semantic tasks of varying complexity: DM_{DE}, a monolingual German SDSM derived from a medium-sized parsed German corpus using simple lexico-syntactic patterns; and the cross-lingual $DM_{EN \rightarrow DE}$, which beyond a source language SDSM requires only a bilingual lexicon. Such DMs are both flexible and powerful: While there is fluctuation in performance of these models across tasks, in general, a number of the various discrete model instances that can be derived from each SDSM – edge, wordsim as well as single- and multi-step prototype models for a given matricization – will achieve high performance and coverage. Beyond the model induction methods, I investigated methods for combining mono- and cross-lingual SDSMs to maximize their performance. Taking the maximum prediction of the models considered was determined to be the most robust and performant combination method. In order to explain why this is a stable effect for distributional methods, I propose the underestimation hypothesis (UEH) which states that noise in a distributional space is more likely to lead to a decrease rather than an increase

in vector similarity. Beyond outlining an argument, I additionally provided empirical evidence in support of UEH. Finally, to test whether the performance patterns found for German, I cross-lingually induced and tested $DM_{EN \rightarrow HR}$ for Croatian and showed that it too could be combined with the monolingual DM_{HR} (Šnajder et al., 2013) to boost accuracy and coverage.

Insights. We found that monolingual models which rely on a target language corpus reliably have higher coverage but with less accurate predictions. Cross-lingual models showed the reverse: Due to their reliance on a bilingual lexicon, the size-limited nature of such resources results in lower coverages on our test scenarios but with higher quality estimates. This led to the investigation of combining mono- and cross-lingual models to ameliorate the negative effects of the single models.

We determined the best model combination scheme to be a late-fusion strategy: Applying a maximum function to the model predictions leads to a high-coverage multilingual model with high-quality predictions. Finally, additional experiments performed during the analysis of the multilingual results indicate that we can further fine-tune our model combination using a frequency threshold on our most accurate model (i.e. our cross-lingual $DM_{EN \rightarrow DE}$/$DM_{EN \rightarrow HR}$) by having them abstain from making predictions on less frequent words.

These patterns of results obtained have been shown to be stable across multiple evaluation scenarios as well as for two quite dissimilar languages. Finally, we suggested the underestimation hypothesis to help explain why maximum performs best, giving evidence of an additive, group-evidence effect on vector dimensions which improves the quality of distributional representations and in our view warrants further experimentation.

Future Work. The most recent and active development in distributional modeling has been the increasing use of neural network methods for building semantic representations called word embeddings. The main difference to the traditional spaces considered in this thesis is that the dimensions in such models are not themselves interpretable but, rather, latent; disallowing such uses such as the duality of a term-document matrix or the various matricizations of an SDSM.

While much work has been invested in developing competing architectures (e.g. formulations of translation matrices and objective functions), it currently appears that the source and type of information used to build cross-lingual embedding models is more important than the specific model architecture which can vary from word-, sentence- or document-aligned data (Ruder, Vulić, and Søgaard, 2017). As possible future work, we would suggest comparing our structured distributional models with some high-performant embedding models. In particular, embedding models that are built from dependency parses (Levy and Goldberg, 2014) but whose representations themselves are not structured could be compared against our distributional models.

A potentially fruitful point of comparison can be found in the popular BERT model (Devlin, Chang, Lee, and Toutanova, 2019) which is built using the attention-based transformer architecture (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin, 2017). It allows for online generation of contextual word representations using any number of languages in pre-training, bringing in multilinguality in a natural way. While our SDSMs go beyond word representations explicitly linking words with e.g. syntactic relations, recent work (Hewitt and Manning, 2019) has shown that such models implicitly encode syntactic information meaning, making them prime candidates for further study.

Finally, in this thesis only model combinations across model types (e.g. XL-, ML-SDSM and DSM) were investigated. However, it is straightforward to conceive of using an SDSM by instantiating many model types simultaneously within a single SDSM and combining those predictions together.

Bibliography

Aina, L. (2014). Dimensionality reduction on syntax-based distributional semantics models: the case of crosslingual and multilingual distributional memories for german. Bachelor's thesis, Università di Pisa.

Anić, V. (2003). *Veliki rječnik hrvatskoga jezika. Dodatak. Pravopisni priručnik: dodatak Velikom rječniku hrvatskoga jezika*. Novi liber.

Baard, M. (2003). AI founder blasts modern research. WIRED.

Baker, C. F., C. J. Fillmore, and J. B. Lowe (1998a). The Berkeley FrameNet Project. In *Proceedings of the joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, Montréal, QC, pp. 86–90.

Baker, C. F., C. J. Fillmore, and J. B. Lowe (1998b). The berkeley framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, Stroudsburg, PA, USA, pp. 86–90. Association for Computational Linguistics.

Baker, C. F., C. J. Fillmore, and J. B. Lowe (1998c). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, Stroudsburg, PA, USA, pp. 86–90. Association for Computational Linguistics.

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta (2008). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation 43*(3), 209–226.

Baroni, M. and A. Kilgarriff (2006). Large Linguistically-Processed Web Corpora

for Multiple Languages. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, pp. 87–90.

Baroni, M. and A. Lenci (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics 36*(4), 1–49.

Baroni, M. and R. Zamparelli (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1183–1193. Association for Computational Linguistics.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology 59*(1), 617–645. PMID: 17705682.

Basili, R. and M. Pennacchiotti (2010). Distributional lexical semantics: Toward uniform representation paradigms for advanced acquisition and processing tasks. *Natural Language Engineering 16*(4), 347–358.

Beckwith, R., C. Fellbaum, D. Gross, and G. A. Miller (1991). Wordnet: A lexical database organized on psycholinguistic principles. *Lexical acquisition: Exploiting on-line resources to build a lexicon*, 211–232.

Beltagy, I., S. Roller, P. Cheng, K. Erk, and R. J. Mooney (2016). Representing meaning with a combination of logical and distributional models. *Computational Linguistics 42*(4), 763–808.

Bentivogli, L. and E. Pianta (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Journal of Natural Language Engineering 11*(3), 247–261.

Blacoe, W. and M. Lapata (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pp. 546–556. Association for Computational Linguistics.

Bohnet, B. (2010). Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, pp. 89–97.

Bos, J. and K. Markert (2005). Recognising textual entailment with logical inference. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, Stroudsburg, PA, USA, pp. 628–635. Association for Computational Linguistics.

Brants, S., S. Dipper, P. Eisenberg, S. Hansen-Schirra, E. König, W. Lezius, C. Rohrer, G. Smith, and H. Uszkoreit (2004, Dec). Tiger: Linguistic interpretation of a german corpus. *Research on Language and Computation 2*(4), 597–620.

Broadbent, D. E. (1967). Word-frequency effect and response bias. *Psychological review 74*(1), 1.

Brockmann, C. and M. Lapata (2003). Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, Stroudsburg, PA, USA, pp. 27–34. Association for Computational Linguistics.

Bruni, E., N.-K. Tran, and M. Baroni (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research 49*, 1–47.

Buchholz, S. and E. Marsi (2006). CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, New York, NY, pp. 149–164.

Budanitsky, A. and G. Hirst (2001). Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and other lexical resources*, Volume 2, pp. 2–2.

Budanitsky, A. and G. Hirst (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics 32*(1).

Chang, H.-S., Z. Wang, L. Vilnis, and A. McCallum (2018). Distributional inclusion vector embedding for unsupervised hypernymy detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 485–495. Association for Computational Linguistics.

Chater, N. and C. D. Manning (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences 10*(7), 335 – 344. Special issue: Probabilistic models of cognition.

Chen, S. F. and J. Goodman (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language 13*(4), 359–394.

Church, K. W. and P. Hanks (1990). Word association norms mutual information, and lexicography. *Computational Linguistics 16*(1).

Clarke, D. (2007). *Context-theoretic Semantics for Natural Language an Algebraic Framework*. Ph. D. thesis, University of Sussex.

Curran, J. R. and M. Moens (2002). Improvements in automatic thesaurus extraction. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, Philadelphia, Pennsylvania, USA, pp. 59–66. Association for Computational Linguistics.

Daelemans, W. and A. Van den Bosch (2005). *Memory-based language processing*. Cambridge University Press.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics.

dict.cc GmbH. dict.cc Online-Wörterbuch.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics 19*(1).

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*, Volume 38. Siam.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. New York, NY: Chapman and Hall.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass 6*(10), 635–653.

Erk, K. and D. McCarthy (2009). Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume*

1 - Volume 1, EMNLP '09, Stroudsburg, PA, USA, pp. 440–449. Association for Computational Linguistics.

Erk, K., S. Padó, and U. Padó (2010). A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics 36*(4), 723–763.

Evert, S. (2005). *The statistics of word cooccurrences: word pairs and collocations*. Ph. D. thesis, IMS, Universität Stuttgart.

Faaß, G. and K. Eckart (2013). Sdewac – a corpus of parsable sentences from the web. In I. Gurevych, C. Biemann, and T. Zesch (Eds.), *Language Processing and Knowledge in the Web*, Volume 8105 of *Lecture Notes in Computer Science*, pp. 61–68. Springer Berlin Heidelberg.

Fano, R. M. (1961). Transmission of information: A statistical theory of communications. *American Journal of Physics 29*(11), 793–794.

Fillmore, C. J. (1976). Frame semantics and the nature of language*. *Annals of the New York Academy of Sciences 280*(1), 20–32.

Firth, J. R. (1957). A synopsis of linguistic theory. *Studies in Linguistic Analysis*, 1–32.

Fischer, G. (2003). *Lineare Gleichungssysteme*. Wiesbaden: Vieweg+Teubner Verlag.

Frege, G. (1884). *Die Grundlagen der Arithmetik: eine logisch mathematische Untersuchung über den Begriff der Zahl*. Verlag von Wilhelm Koebner.

Fried, D., T. Polajnar, and S. Clark (2015). Low-rank tensors for verbs in compositional distributional semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 731–736. Association for Computational Linguistics.

Fung, P. and B. Chen (2004). BiFrameNet: Bilingual Frame Semantics Resources Construction by crosslingual Induction. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, pp. 931–935.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. Technical Report 257, University of Illinois at Urbana-Champaign - Center for the Study of Reading.

Gentner, D. (2006). Why verbs are hard to learn. *Action meets word: How children learn verbs*, 544–564.

Grefenstette, G. (1994a). *Explorations in Automatic Thesaurus Discovery*. Kluwer.

Grefenstette, G. (1994b). *Explorations in Automatic Thesaurus Discovery*. Boston/Norwell, MA: Kluwer Academic Publishers.

Guevara, E. (2010). A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pp. 33–37. Association for Computational Linguistics.

Gunes, H. and M. Piccardi (2005, Oct). Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, Volume 4, pp. 3437–3443 Vol. 4.

Gurevych, I. (2005). Using the structure of a conceptual network in computing semantic relatedness. In *Second International Joint Conference on Natural Language Processing: Full Papers*.

Haghighi, A., P. Liang, T. Berg-Kirkpatrick, and D. Klein (2008, June). Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, Columbus, Ohio, pp. 771–779. Association for Computational Linguistics.

Halácsy, P., A. Kornai, and C. Oravecz (2007). Hunpos: An open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, Stroudsburg, PA, USA, pp. 209–212. Association for Computational Linguistics.

Hamp, B. and H. Feldweg (1997). Germanet - a lexical-semantic net for german. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Harris, Z. S. (1954). Distributional structure. *Word 10*(23), 146–162.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational linguistics*, Volume 2, pp. 539–545. Association for Computational Linguistics.

Henrich, V. and E. Hinrichs (2010). Gernedit - the germanet editing tool. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA).

Hewitt, J. and C. D. Manning (2019, June). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, pp. 4129–4138. Association for Computational Linguistics.

Hwa, R., P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak (2005). Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Journal of Natural Language Engineering 11*(3), 311–325.

Ingason, A. K., S. Helgadóttir, H. Loftsson, and E. Rögnvaldsson (2008). A mixed method lemmatization algorithm using a hierarchy of linguistic identities (holi). In B. Nordström and A. Ranta (Eds.), *Advances in Natural Language Processing*, Berlin, Heidelberg, pp. 205–216. Springer Berlin Heidelberg.

J. Jiang, J. and D. W. Conrath (1997, 10). Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of the International Conference on Research in Computational Linguistics 10*.

Janković, V., J. Šnajder, and B. D. Bašić (2011). Random indexing distributional semantic models for croatian language. In *Text, Speech and Dialogue*, pp. 411–418. Springer.

Joanis, E., S. Stevenson, and D. James (2006). A general feature space for automatic verb classification. *Natural Language Engineering 14*(03), 337–367.

Kamp, H. and U. Reyle (1993). *From Discourse to Logic: Introduction to modeltheoretic semantics of natural language, formal logic and Discourse Representation Theory*. Kluwer Academic Publishers.

Karan, M., J. Šnajder, and B. Dalbelo Bašić (2012). Distributional semantics approach to detecting synonyms in Croatian language. In *Proceedings of the Eighth Language Technologies Conference*, Ljubljana, Slovenia.

Kartsaklis, D. (2015). *Compositional distributional semantics with compact closed categories and frobenius algebras*. Ph. D. thesis, University of Oxford.

Kartsaklis, D. and M. Sadrzadeh (2016). Distributional inclusion hypothesis for tensor-based composition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2849–2860. The COLING 2016 Organizing Committee.

Kiela, D. and S. Clark (2015, September). Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 2461–2470. Association for Computational Linguistics.

Kiela, D. and S. Clark (2017). Learning neural audio embeddings for grounding semantics in auditory perception. *J. Artif. Intell. Res. 60*, 1003–1030.

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM review 51*(3), 455–500.

Kotlerman, L., I. Dagan, I. Szpektor, and M. Zhitomirsky-Geffet (2009). Directional distributional similarity for lexical expansion. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 69–72. Association for Computational Linguistics.

Lakoff, G. and M. Johnson (1980). *Metaphors we live by*. University of Chicago press.

Lan, Y. and J. Jiang (2018, August). Embedding wordnet knowledge for textual entailment. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, pp. 270–281. Association for Computational Linguistics.

Landauer, T. K. and S. T. Dumais (1997a). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211–240.

Landauer, T. K. and S. T. Dumais (1997b). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104(2)*, 211–240.

Landauer, T. K., D. S. McNamara, S. E. Dennis, and W. E. Kintsch (2007). Handbook of latent semantic analysis.

Lapata, M., F. Keller, and C. Scheepers (2003). Intra-sentential context effects on the interpretation of logical metonymy. *Cognitive Science 27*(4), 649–668.

Lapata, M. and A. Lascarides (2003). A probabilistic account of logical metonymy. *Computational Linguistics 29*(2), 263–317.

Lapesa, G., S. Evert, and S. Schulte im Walde (2014). Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, Dublin, Ireland, pp. 160–170. Association for Computational Linguistics and Dublin City University.

Lenat, D. B., R. V. Guha, K. Pittman, D. Pratt, and M. Shepherd (1990, August). Cyc: Toward programs with common sense. *Commun. ACM 33*(8), 30–49.

Lenci, A. (2011). Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, Portland, Oregon, pp. 58–66.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pp. 24–26. ACM.

Levy, O. and Y. Goldberg (2014, June). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, Maryland, pp. 302–308. Association for Computational Linguistics.

Lewis, M. and M. Steedman (2013). Combined distributional and logical semantics. *Transactions of the Association for Computational Linguistics 1*, 179–192.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, Montreal, QC, pp. 768–774.

Lin, D. et al. (1998). An information-theoretic definition of similarity. In *Icml*, Volume 98, pp. 296–304. Citeseer.

Ljubešić, N. and T. Erjavec (2011). hrwac and slwac: Compiling web corpora for croatian and slovene. In I. Habernal and V. Matoušek (Eds.), *Text, Speech and Dialogue*, Berlin, Heidelberg, pp. 395–402. Springer Berlin Heidelberg.

Lowe, W. (2001). Towards a theory of semantic space. In J. D. Moore and K. Stenning (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society*, Volume 23, pp. 576–581.

Maedche, A. and S. Staab (2000). Mining ontologies from text. In R. Dieng and O. Corby (Eds.), *Knowledge Engineering and Knowledge Management Methods, Models, and Tools*, Berlin, Heidelberg, pp. 189–202. Springer Berlin Heidelberg.

McCarthy, D., R. Koeling, J. Weeds, and J. Carroll (2004). Finding Predominant Word Senses in Untagged Text. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, pp. 279–286.

McDonald, R., K. Lerman, and F. Pereira (2006, June). Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, New York City, pp. 216–220. Association for Computational Linguistics.

McDonald, R., J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Täckström, C. Bedini, N. Bertomeu Castelló, and J. Lee (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, pp. 92–97.

McDonald, R., F. Pereira, K. Ribarov, and J. Hajič (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, BC, pp. 523–530.

McRae, K., T. R. Ferretti, and L. Amyote (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes 12*(2-3), 137–176.

McRae, K., M. Hare, J. L. Elman, and T. Ferretti (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition 33*(7), 1174–1184.

McRae, K., M. J. Spivey-Knowlton, and M. K. Tanenhaus (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language 38*(3), 283–312.

Merlo, P. and S. Stevenson (2001). Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics 27*(3), 373–408.

Michelbacher, L., S. Evert, and H. Schütze (2007). Asymmetric association measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2007*.

Mihalcea, R. and D. Radev (2011). *Graph-based Natural Language Processing and Information Retrieval*. Cambridge, UK: Cambridge University Press.

Miller, G. A. (1986). Dictionaries of the mind. *Language and Cognitive Processes*, 171–185.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller (1990). Introduction to wordnet: An on-line lexical database*. *International Journal of Lexicography 3*(4), 235–244.

Miller, G. A. and C. Fellbaum (1992). Wordnet and the organization of lexical memory. In M. L. Swartz and M. Yazdani (Eds.), *Intelligent Tutoring Systems for Foreign Language Learning*, Berlin, Heidelberg, pp. 89–102. Springer Berlin Heidelberg.

Mitchell, J. and M. Lapata (2010). Composition in distributional models of semantics. *Cognitive Science 34*(8), 1388–1429.

Mitchell, J., M. Lapata, V. Demberg, and F. Keller (2010). Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Stroudsburg, PA, USA, pp. 196–206. Association for Computational Linguistics.

Mohammad, S., I. Gurevych, G. Hirst, and T. Zesch (2007). Crosslingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic, pp. 571–580.

Montague, R. (1970). Universal grammar. *Theoria 36*(3), 373–398.

Navigli, R. and S. P. Ponzetto (2012, December). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell. 193*, 217–250.

Navigli, R. and P. Velardi (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Stroudsburg, PA, USA, pp. 1318–1327. Association for Computational Linguistics.

Nivre, J. (2006). *Inductive Dependency Parsing*. Dordrecht, Netherlands: Springer.

Nivre, J. (2008). Algorithms for deterministic incremental dependency parsing. *Computational Linguistics 34*(4), 513–553.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General 115*(1), 39.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. *Formal approaches in categorization*, 18–39.

Padó, S. and M. Lapata (2007). Dependency-based construction of semantic space models. *Computational Linguistics 33*(2), 161–199.

Padó, S., J. Šnajder, and B. Zeller (2013). Derivational smoothing for syntactic distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 731–735. Association for Computational Linguistics.

Padó, S., J. Šnajder, J. Utt, and B. Zeller (2016). Smoothing syntax-based semantic spaces: Let the winner take it all. In *Proceedings of KONVENS*, Bochum, Germany, pp. 186–191. Acceptance rate: 65%.

Palermo, D. S. and J. J. Jenkins (1964). Word association norms: Grade school through college.

Panther, K.-U. and G. Radden (1999). *Metonymy in language and thought*, Volume 4. John Benjamins Publishing.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Patel, M., J. A. Bullinaria, and J. P. Levy (1998). Extracting semantic representations from large text corpora. In J. A. Bullinaria, D. W. Glasspool, and G. Houghton (Eds.), *4th Neural Computation and Psychology Workshop*, London, pp. 199–212. Springer London.

Peirsman, Y., K. Heylen, and D. Geeraerts (2008). Size matters: tight and loose context definitions in english word space models. In *Proceedings of the ESSLLI workshop on distributional lexical semantics*, pp. 34–41. Springer.

Peirsman, Y. and S. Padó (2011). Semantic relations in bilingual lexicons. *ACM Transactions in Speech and Language Processing 8*(2), 3:1–3:21.

Pennington, J., R. Socher, and C. Manning (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543. Association for Computational Linguistics.

Prescher, D., S. Riezler, and M. Rooth (2000). Using a Probabilistic Class-Based Lexicon for Lexical Ambiguity Resolution. In *Proceedings of COLING 2000*, Saarbrücken, Germany.

Rada, R., H. Mili, E. Bicknell, and M. Blettner (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics 19*(1), 17–30.

Rapp, R. (2009). The backtranslation score: Automatic mt evalution at the sentence level without reference translations. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Suntec, Singapore, pp. 133–136. Association for Computational Linguistics.

Resnik, P. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph. D. thesis, University of Pennsylvania.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, QC, Canada, pp. 448–453.

Roller, S. and K. Erk (2016). Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment. In *Proceedings*

of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2163–2172. Association for Computational Linguistics.

Roller, S., K. Erk, and G. Boleda (2014). Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1025–1036. Dublin City University and Association for Computational Linguistics.

Rothenhäusler, K. and H. Schütze (2009, March). Unsupervised classification with dependency based word spaces. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, Athens, Greece, pp. 17–24. Association for Computational Linguistics.

Ruder, S., I. Vulić, and A. Søgaard (2017, Jun). A Survey Of Cross-lingual Word Embedding Models. *arXiv e-prints*, arXiv:1706.04902.

Sadat, F., M. Yoshikawa, and S. Uemura (2003). Learning bilingual translations from comparable corpora to cross-language information retrieval: hybrid statistics-based and linguistics-based approach. In *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages*, Sapporo, Japan, pp. 57–64.

Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph. D. thesis.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Salton, G. and C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24*(5), 513 – 523.

Sayeed, A., V. Demberg, and P. Shkadzko (2015). An exploration of semantic features in an unsupervised thematic fit evaluation framework. *Italian Journal of Computational Linguistics 1*(1).

Schulte im Walde, S. (2006). Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics 32*(2), 159–194.

Schütze, H. (1993). Part-of-speech induction from scratch. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, USA, pp. 251–258. Association for Computational Linguistics.

Shwartz, V., Y. Goldberg, and I. Dagan (2016). Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2389–2398. Association for Computational Linguistics.

Šnajder, J., S. Padó, and v. Agić (2013). Building and evaluating a distributional memory for Croatian. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, pp. 784–789.

Snoek, C. G. M., M. Worring, and A. W. M. Smeulders (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, New York, NY, USA, pp. 399–402. ACM.

Soderland, S., O. Etzioni, D. S. Weld, M. Skinner, and J. Bilmes (2009). Compiling a Massive, Multilingual Dictionary via Probabilistic Inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, Suntec, Singapore, pp. 262–270.

Somers, H. (2005). Round-trip translation: What is it good for? In *Proceedings of the Australasian Language Technology Workshop*, Sydney, Australia, pp. 127–133.

Szpektor, I. and I. Dagan (2008). Learning entailment rules for unary templates. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, Stroudsburg, PA, USA, pp. 849–856. Association for Computational Linguistics.

Taft, M. (1979, Jul). Recognition of affixed words and the word frequency effect. *Memory & Cognition 7*(4), 263–272.

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics 32*(3), 379–416.

Turney, P. D. and P. Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research 37*, 141–188.

Utt, J. and S. Padó (2014). Crosslingual and multilingual construction of syntax-based vector space models. *Transactions of the Association for Computational Linguistics 2*, 245–258.

van Gompel, R. P., M. J. Pickering, J. Pearson, and G. Jacob (2006). The activation of inappropriate analyses in garden-path sentences: Evidence from structural priming. *Journal of Memory and Language 55*(3), 335 – 362.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc.

Wallace, D. and L. A. Wallace (2005). *Reader's Digest, das Beste für Deutschland*. Stuttgart, Germany: Verlag Das Beste.

Weeds, J., D. Weir, and D. McCarthy (2004). Characterising measures of lexical distributional similarity. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*.

Westerhout, E. (2009). Definition extraction using linguistic and structural features. In *Proceedings of the 1st Workshop on Definition Extraction*, WDE '09, Stroudsburg, PA, USA, pp. 61–67. Association for Computational Linguistics.

Wild, F., C. Stahl, G. Stermsek, and G. Neumann (2008). Parameters driving effectiveness of automated essay scoring with LSA. In *Proceedings of the 9th Computer-Aided Assessment Conference*, Loughborough, UK, pp. 485–494.

Winograd, T. (1978). On primitives, prototypes, and other semantic anomalies. In *Proceedings of the 1978 Workshop on Theoretical Issues in Natural Language Processing*, TINLAP '78, Stroudsburg, PA, USA, pp. 25–32. Association for Computational Linguistics.

Wittgenstein, L. (1953). *Philosophische Untersuchungen*. Blackwell. Translated into English by G.E.M. Anscombe.

Wu, S. and W. Schuler (2011). Structured composition of semantic vectors. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Xia, F., W. D. Lewis, M. W. Goodman, G. Slayden, R. Georgi, J. Crowgey, and E. M. Bender (2016, Jun). Enriching a massively multilingual database of interlinear glossed text. *Language Resources and Evaluation 50*(2), 321–349.

Yarowsky, D. and G. Ngai (2001). Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora. In *Proceedings of the 2nd Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA, pp. 200–207.

Zarcone, A., A. Lenci, S. Padó, and J. Utt (2013). Fitting, not clashing! a distributional semantic model of logical metonymy. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, Potsdam, Germany, pp. 404–410. Association for Computational Linguistics.

Zarcone, A. and S. Padó (2011). Generalized event knowledge in logical metonymy resolution. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, Boston, Massachusetts, pp. 944–949.

Zarcone, A., S. Padó, and A. Lenci (2012). Inferring covert events in logical metonymies: a probe recognition experiment. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, Sapporo, Japan.

Zarcone, A., J. Utt, and S. Padó (2012). Modeling covert event retrieval in logical metonymy: Probabilistic and distributional accounts. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics*, CMCL '12, Stroudsburg, PA, USA, pp. 70–79. Association for Computational Linguistics.

Zeller, B., J. Šnajder, and S. Padó (2013). DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of ACL 2013*, Sofia, Bulgaria, pp. 1201–1211.

Zesch, T., I. Gurevych, and M. Mühlhäuser (2007). Comparing Wikipedia and German Wordnet by evaluating semantic relatedness on multiple datasets. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, pp. 205–208.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press, Inc.

Soli Deo gloria.

Ecclesia semper reformanda est.