Crosslingual and Multilingual Construction of Syntax-Based Vector Space Models

Jason Utt and **Sebastian Padó**
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
[uttjn|pado]@ims.uni-stuttgart.de

Abstract

Syntax-based distributional models of lexical semantics provide a flexible and linguistically adequate representation of co-occurrence information. However, their construction requires large, accurately parsed corpora, which are unavailable for most languages.

In this paper, we develop a number of methods to overcome this obstacle. We describe (a) a *crosslingual* approach that constructs a syntax-based model for a new language requiring only an English resource and a translation lexicon; and (b) *multilingual* approaches that combine crosslingual with monolingual information, subject to availability. We evaluate on two lexical semantic benchmarks in German and Croatian. We find that the models exhibit complementary profiles: crosslingual models yield higher accuracies while monolingual models provide better coverage. In addition, we show that simple multilingual models can successfully combine their strengths.

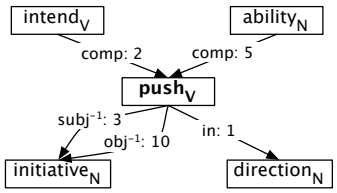
1 Introduction

Building on the Distributional Hypothesis (Harris, 1954; Miller and Charles, 1991), which states that words occurring in similar contexts are similar in meaning, distributional semantic models (DSMs) represent a word's meaning via its occurrence in context in large corpora. Vector spaces, the most widely used type of DSMs, represent words as vectors in a high-dimensional space whose dimensions correspond to features of the words' contexts. *Word spaces* represent the simplest case of DSMs in which the dimensions are simply the context words (Schütze, 1992). A notable subclass of DSMs are *syntax-based models* (Lin, 1998; Baroni and Lenci, 2010) which use (lexicalized) syntactic relations as dimensions. They are able to model more fine-grained distinctions than word spaces and have been found to be useful for tasks such as selectional preference learning (Erk et al., 2010), verb class induction (Schulte im Walde, 2006), analogical reasoning (Turney, 2006), and alternation discovery (Joanis et al., 2006). Despite their flexibility and usefulness, syntax-based DSMs are used less often than word-based spaces. An important reason is that their construction requires accurate parsers, which are unavailable for many languages. In addition, syntax-based DSMs are inherently more sparse than word spaces, which calls for a large corpus of well parsable data. It is thus not surprising that besides English (Baroni and Lenci, 2010), only few other languages possess large-scale syntax-based DSMs (Padó and Utt, 2012; Šnajder et al., 2013).

This paper develops methods that take advantage of the *resource gradient* between English and other languages, exploiting the higher-quality resources of the former to induce resources for target languages among the latter, by *translating* the *word-link-word* co-occurrences that underlie syntax-based DSMs. This directly provides a *crosslingual method* to construct syntax-based DSMs for target languages without any target language data, requiring only an English syntax-based DSM and a translation lexicon. Such lexicons are available for many language pairs, and we outline a method to reduce ambiguity inherent in such dictionaries. We describe a set of *multilingual methods* that can combine corpus evidence from English and the target language to further improve the performance of the obtained DSM.

We consider two target languages, German and Croatian, as examples of one close and one more remote target language. For evaluation, we use two

Transactions of the Association for Computational Linguistics, 2 (2014) 245–258. Action Editor: Patrick Pantel.
Submitted 11/2013; Revised 3/2014; Published 10/2014.



(a) DM as graph

$W \times LW$	push
$\langle \text{comp}^{-1}, \text{intend} \rangle$	2
$\langle \text{comp}^{-1}, \text{ability} \rangle$	5
$\langle \text{subj}^{-1}, \text{initiative} \rangle$	3
$\langle \text{obj}^{-1}, \text{initiative} \rangle$	10
$\langle \text{in}, \text{direction} \rangle$	1

(b) DM as $W \times LW$ matrix

$WL \times W$	$\langle$push, comp$^{-1}\rangle$	$\langle$push, in$\rangle$	$\langle$push, obj$^{-1}\rangle$	$\langle$push, subj$^{-1}\rangle$
ability	5	0	0	0
direction	0	1	0	0
initiative	0	0	10	3
intend	2	0	0	0

(c) DM as $WL \times W$ matrix

Figure 1: Distributional Memory sample around *to push* represented as a graph (a) and two matrices (b, c)

tasks, namely synonym choice and semantic similarity prediction. For both languages and tasks, monolingually constructed DSMs can provide strong baselines. We find similar patterns across tasks and target languages: the crosslingually constructed DSM can be parametrized so that it becomes superior to an existing monolingual DSM in quality, even if inferior in coverage. A simple multilingual backoff can combine the crosslingual model's high quality with the monolingual model's high coverage.

Structure of the paper. We begin by sketching the structure of Distributional Memory, a general framework for syntax-based semantic spaces, in Section 2. Our main contributions follow in Sections 3 and 4, namely, a family of models for the crosslingual and multilingual construction of DSMs. The second part of the paper is concerned with evaluation. Section 5 describes our experimental setup after which we discuss our results for German (Section 6) and Croatian (Section 7). The paper concludes with related work (Section 8) and a general discussion (Section 9).

2 Distributional Memory: A General Model of Syntax-based Vector Spaces

2.1 Motivation and Definition

Simple syntax-based DSMs represent target words in terms of dimensions labeled with word-relation pairs (Lin, 1998; Grefenstette, 1994). Unfortunately, this representation only supports tasks that compare pairs of words with regard to their meaning (e.g., in synonymy detection or selectional preferences), but not for tasks such as analogical reasoning, where sets of word pairs are compared (Turney, 2006).

To unify syntax-based DSMs, Baroni and Lenci (2010) proposed the *Distributional Memory* (DM) model which captures distributional information at the more general level of word-link-word triples, stored as a third order co-occurrence tensor. The DM tensor can be seen as a set of ordered *word-link-word* tuples such as $\langle$*pencil obj use*$\rangle$ associated with a scoring function $\sigma\colon W \times L \times W \to \mathbb{R}^+$ that scores, for example, $\langle$*pencil obj use*$\rangle$ more highly than $\langle$*elephant obj use*$\rangle$.

The DM tensor can be visualized as a directed graph whose nodes are labeled with lemmas and whose edges are labeled with links and scores. As an example, Figure (1a) shows five links for the verb *push* in the English DM, including subject, object, prepositional adjunct, and governing verbs.

DSMs for individual tasks can be obtained by "matricizing" the tensor into two-dimensional matrices corresponding to standard vector spaces. The matrix in Figure (1b) shows the *word by link-word* space ($W \times LW$). It represents words w in terms of pairs $\langle l, w \rangle$ of a link and a context word. This space models similarity among words, e.g. for thesaurus construction (Lin, 1998). The example matrix in Figure (1c) represents a *word-link by word* space ($WL \times W$). It characterizes pairs $\langle w, l \rangle$ through context words w, which can be understood as selectional preferences.

DM does not assume a specific source for building the graph. However, all existing DM resources were extracted from large dependency-parsed corpora such as UKWAC (Baroni et al., 2008). In the simplest case, the set of labels L is (a subset of) the dependency relations in the corpus, and the scoring function σ is a measure of association between the governor and the dependent (see Baroni and Lenci (2010) for details). However, the most robust DMs (including Baroni and Lenci's LexDM and TypeDM) use both syntactic and lexicalized links, i.e. links which contain words themselves, as well as *surface form-based* links, e.g., observed *subject-verb-object* triples in the corpus lead to a $\langle$*subject verb object*$\rangle$ edge in the DM graph.

2.2 DMs for Other Languages

Given the appealing properties of Distributional Memory, it may be surprising that not many comparable resources exist for other languages. To our knowledge, comparable resources exist only for German (Padó and Utt, 2012) and Croatian (Šnajder et al., 2013). Both studies replicate the monolingual DM construction process outlined by Baroni and Lenci for the respective languages. For German, the process is relatively unproblematic, since German is relatively well-equipped in terms of corpora and parsers. In contrast, Šnajder et al. (2013) faced serious resource scarcity while building a Croatian DM and had to go to considerable lengths to clean a large web corpus and to optimize the linguistic processing tools. The resulting DM outperforms a monolingual context word model for nouns and verbs, but performs worse than the word-based model for (generally rarer) adjectives. As a direct consequence, high-quality syntax-based DSMs can only be constructed for a limited set of languages.

3 Crosslingual Construction of DMs

3.1 Motivation

As outlined in the previous section, there is a bottleneck in many languages regarding both large, clean corpora as well as processing pipelines that result in high-accuracy dependency parses. To address this problem, we propose to induce Distributional Memories for such languages crosslingually by *translating* a source language DM into the target language.

By adopting English as the source language we can take advantage of the *resource gradient*, that is, the higher maturity of English NLP techniques, such as parsers, compared to most other languages. For many languages, treebanks have become available only within the last ten years (Buchholz and Marsi, 2006), if at all, while English has been at the forefront of NLP development for several decades, and a number of highly accurate dependency parsers exist (McDonald et al., 2005; Nivre, 2006). At the same time, English arguably possesses the widest range of large and well-cleaned corpora of any language.

To make our approaches applicable to as many target languages as possible, we assume in this section that very few resources for the target language are available. The crosslingual methods we develop here work without any target language corpora, either monolingual or bilingual. The only knowledge we use is a simple *translation lexicon*, that is, a list of translation pairs without translation probabilities, as shown in Figure 2. Translation lexicons of this type are arguably the most common bilingual resource and accurate ones exist for virtually any language pair (Soderland et al., 2009), even for languages with few available corpora. Furthermore, such translation lexicons are often crowdsourced and are available for download. For example, the website `dict.cc` provides numerous such lexicons for German and English.

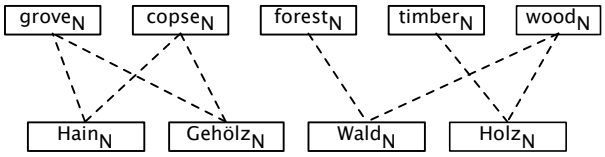


Figure 2: Sample of the English-German `dict.cc` dictionary; translations shown as dashed lines.

This approach promises in particular to yield models with a quality-coverage profile complementary to that of monolingual models (Mohammad et al., 2007; Peirsman and Padó, 2011): Crosslingual DMs are extracted from source language corpora which we assume to be parsed more accurately than target language corpora. In addition, the translation process can be designed to act as a further filtering step (cf. Section 3.4 below), thus optimizing crosslingual models for higher quality at the expense of coverage. In contrast, monolingual models – in particular for under-resourced languages – often hit a quality ceiling, but can generally guarantee high coverage.

3.2 Translating DMs with Translation Lexicons

We conceptualize DM as a directed graph (see Figure 1), which allows us to phrase translation in graph terms (Mihalcea and Radev, 2011). A DM is a triple (V, E, σ) where V is a set of vertices (i.e., the vocabulary), E a set of typed edges between words, represented as word-link-word triples (cf. Section 2), and σ, an edge-weighting function. We will use S and T to refer to the source and target language vocabularies, respectively, and (V_S, E_S, σ_S) and (V_T, E_T, σ_T) to denote source and target language DMs.

We can now ask how the shape of the graph

changes under translation. In an ideal world, a translation lexicon would be a bijective function between the source and target language vocabularies: $\mathrm{Tr} : S \to T$. Then, the transformation would merely constitute a relabeling. We would then construct the German DM graph by exchanging all English node labels with German node labels, i.e., $V_T = T$, and creating a German edge for each English edge.[1]

3.3 Ambiguity in Unfiltered Translation

The dictionary fragment in Figure 2 shows that translation is not bijective but a many-to-many relation. In fact, taking the English–German `dict.cc` lexicon as an example, there is an average of 2.3 German translations for each English lemma, and an average of 1.9 English translations for each German lemma. We model this situation using two functions: $\mathrm{Tr} : S \to 2^T$ translates source words into sets of target words, and $\mathrm{Tr}^{-1} : T \to 2^S$ translates target words back into the source language.

The naive way to translate nodes using Tr is to use *all* translations for a given word. Thus, for each edge in the source DM between lemmas s_1 and s_2, we obtain $|\,\mathrm{Tr}(s_1)| \cdot |\,\mathrm{Tr}(s_2)|$ edges in the target language:

$$E_T = \{(t_1, l, t_2) \,|\, \exists (s_1, l, s_2) \in E_S : t_1 \in \mathrm{Tr}(s_1) \wedge t_2 \in \mathrm{Tr}(s_2)\} \quad (1)$$

The score σ_T of a target edge is defined as the mean of the scores of all source edges that map to it.

$$\sigma_T(t_1, l, t_2) = \sum_{\substack{s_1 \in \mathrm{Tr}^{-1}(t_1) \\ s_2 \in \mathrm{Tr}^{-1}(t_2)}} \frac{\sigma_S(s_1, l, s_2)}{|\,\mathrm{Tr}^{-1}(t_1)| \cdot |\,\mathrm{Tr}^{-1}(t_2)|} \quad (2)$$

We take the mean as it is less sensitive to outliers than maximum or minimum. In addition, unlike taking the sum, it is also automatically normalized regarding the number of translations, thus penalizing words with many unrelated senses.

A look at Figures 1 and 3, however, indicates that this procedure overgenerates. This is problematic on two levels. First, the target language graph will contain a very large number of edges (e.g., using `dict.cc`, the edge ⟨*text sbj_tr use*⟩ has 42 German translations). Second, the correctness of the target DM suffers. For some cases, such as *copse – Gehölz, Hain*, the various translations are synonymous, and Eq. (1) is appropriate. In other cases, multiple translations indicate *lexical ambiguity* of the source term. For example, the two translations of *wood* correspond to its senses as *forest (Wald)* and *timber (Holz)*, respectively. In such cases, Eq. (1) confuses the senses, as the example in Figure 3 illustrates. The left-hand side shows DM edges between *wood* and two adjectival modifiers, namely *precut* (which is more plausible for the *timber* sense) and *great* (which is more plausible for the *forest* sense). The right-hand side shows (part of) the German translations according to Eq. (1): both *Holz (timber)* and *Wald (forest)* are linked to both adjectives, leading to spurious edges in the German DM.

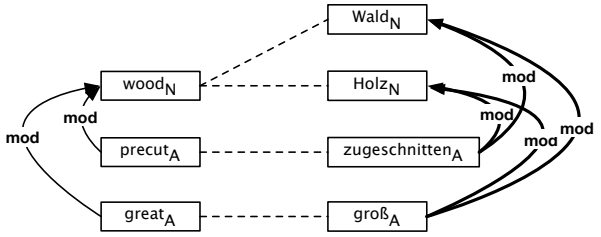


Figure 3: Unfiltered edge translation (EN–DE)

3.4 Filtering by Backtranslation

Since the nature of the translation is not indicated in the translation lexicon, we exploit typical redundancies in the source DM, which often contains "quasi-synonymous" edges that express the same relation with different words, e.g., ⟨*book obj read*⟩ and ⟨*novel obj read*⟩. This allows us to score target edge candidates by how well we can "backtranslate" (Somers, 2005) them into the source language.

This idea is illustrated in Figure 4. We still assume, as above, that *wood* has two translations, but that *precut* has only one. For the English edge ⟨*precut mod wood*⟩, we obtain two German candidate edges, namely ⟨*zugeschnitten mod Holz*⟩ and ⟨*zugeschnitten mod Wald*⟩. When backtranslating these candidates, the first one, ⟨*zugeschnitten mod Wald*⟩, maps only onto the original edge. The second one, ⟨*zugeschnitten mod Holz*⟩, is backtranslated into a different source edge, ⟨*precut mod timber*⟩, which makes it more probable.

[1] We build on the assumption that dependency relations are language-independent which, while incorrect, represents a reasonable simplification (McDonald et al., 2013).

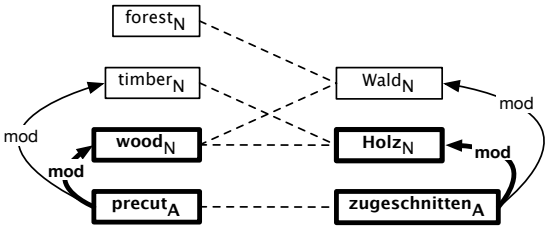


Figure 4: Backtranslation filtering. Original and winning edges shown in boldface.

Model	Covered items Corr.	Cov.
DM.DE (AllL)	.43	.60
DM.DE (SPrfL)	.43	.60
DM.XL EN→DE filter (AllL)	.42	**.61**
DM.XL EN→DE filter (SPrfL)	**.49**	.49

Table 1: Coverage and Correlation (Pearson's r) for predicting word similarity, contrasting link types (all links vs. selectional preference links)

We operationalize this by adding another condition to Eq. (1), namely that target edges must be among the highest-scoring edges for some source edge. Recall that our target scores σ_T are already defined in terms of source edge scores, so no redefinition of the scoring function is necessary.

$$\begin{aligned} E_T = \{(t_1, l, t_2) \,|\, \exists\, (s_1, l, s_2) \in E_S : \\ t_1 \in \mathrm{Tr}(s_1) \wedge t_2 \in \mathrm{Tr}(s_2) \wedge \\ \sigma_T(t_1, l, t_2) = \max_{\substack{t \in \mathrm{Tr}(s_1) \\ t' \in \mathrm{Tr}(s_2)}} \sigma_T(t, l, t'), \} \end{aligned} \quad (3)$$

where $\sigma_T(t, l, t')$ is the score as defined in Eq. 2. This filtering scheme is fairly liberal: we do not limit the number of target edges that a source edge can translate to. A stricter variant could, e.g., abstain from translating a source edge if no unique best edge exists. We leave such variants to future work.

3.5 Defining Similarity

Recall from Figure 1 that DM contains information about both "incoming" as well as "outgoing" links. Monolingually constructed DMs by default use all of these relations since the information is reliable. The situation is not as clear in a crosslingual setting. Our intuition is that *selectional preferences* are most informative and most likely to survive translation. For example, for verbs we expect knowledge about their arguments to be more informative than about their governors. Conversely, for nouns we want to use knowledge about the verbs that they occur with rather than their arguments or modifiers.

We implement this idea by computing semantic similarity between word vectors either on complete vectors (condition "AllL") or on a filtered version that uses only inverse links for verbs and only regular links for nouns and adjectives (condition "SPrfL"). Table 1 shows the results of preliminary experiments on a semantic similarity dataset (details in Section 5). They bear out our hypothesis: in the monolingual setting, there is almost no difference. Thus, in line with previous work, we adopt (AllL) for DM.DE. In contrast, we see a clear quality-coverage trade-off in the crosslingual scenario, with a higher quality for (SPrfL). Since this corresponds to our focus on higher precision for crosslingual models, we will adopt (SPrfL) for all crosslingual DMs.

4 Multilingual Construction of DMs

The crosslingual models described in the previous section do not use any corpus information from the target language: As previously discussed, our rationale is to make the methods as widely applicable as possible. However, this assumption may be too cautious as more corpora and parsers continually become available. In order to take advantage of such developments, this section discusses two simple methods for combining monolingually and crosslingually constructed DMs, thereby combining corpus evidence from both the source and the target language.

We concentrate on methods that can be applied to DMs directly, e.g. by researchers who do not have access to the source corpora. Moreover, we combine not the graphs, but the resulting semantic similarities.[2] We take our inspiration from work on combining and smoothing n-gram language models, where the usual operations are interpolation and back-off (Chen and Goodman, 1998). Note that in our case, the two models to be combined are assumed to have complementary properties, with the monolin-

[2] We conducted experiments with graph merging but found that the different topologies of the monolingual and crosslingual DMs make it difficult to merge the graphs in a manner that combines the information from both graphs.

gual model having higher coverage and the crosslingual model higher quality (cf. Section 3.1). For this reason, we assume that a linear interpolation of the models' similarities for each word pair will not perform well. Our first strategy is a simple backoff combination (DM.MULTI Backoff) that starts with the crosslingual model and falls back to the monolingual model in the case of zero-similarities. Our second strategy follows the intuition that both noise and sparse data tend to result in underestimated similarities. This leads us to the DM.MULTI MaxSim model: It takes the predictions from the monolingual and crosslingual model and takes the higher one.

Both DM.MULTI variants combine predictions from two models and implicitly assume that the predictions are drawn from the same score distribution. Since this is not guaranteed, we standardize all scores before combination, that is, we linearly transform the values so that the resulting distribution has a mean of 0 and a standard deviation of 1.

5 Experimental Setup

To show the benefits of our crosslingual methods, we perform experiments for the language pairs English–German and English–Croatian. These languages exemplify variability on the resource gradient: The resource situation is best for English, still relatively good for German, and most difficult for Croatian.

This section outlines the experiments for German; Section 7 focuses on Croatian. We evaluate our models on two standard tasks from lexical semantics: synonym choice and the prediction of human relatedness judgments. Even though these two tasks are *in-vitro*, they are widely used for model selection in distributional space models and we can compare the results of our models against previous work. The two tasks test how well the models can account for two different aspects of lexical semantics, namely a specific lexical relation (synonymy) and general semantic relatedness.

5.1 Tasks and Datasets

Our first task is synonym detection, where models have to identify the true synonym for a target word from four candidates. We use the German Reader's Digest Word Power (RDWP) dataset (Wallace and Wallace, 2005) which contains 984 items.[3] RDWP is similar to the English TOEFL data (Landauer and Dumais, 1997), but can contain short phrases among the candidates (cf. example in Table 2a).

	Demagoge		**demagogue**
1	*Miesmacher*	×	*grinch*
2	*guter Redner*	×	*able speaker*
3	*skrupelloser Hetzer*	✓	*unscrupulous agitator*
4	*Meinungsforscher*	×	*pollster*

(a) Task 1: synonym target with four candidates

Word Pair	Similarity
Absage - ablehnen *(rejection - refuse)*	3.5
Absage - Stellenanzeige *(rejection - job advertisement)*	1.875
Affe - Gepäckkontrolle *(monkey - luggage inspection)*	0.125

(b) Task 2: semantic similarity (range: 0–4)

Table 2: Example items from evaluation tasks

Our second evaluation tests how well the models predict similarities for German word pairs including closely related, somewhat related, and unrelated word pairs (cf. Table 2b). We use the Gur350 dataset[4] which contains 350 word pairs scored for relatedness by native German taggers on a five-point Likert scale between 0 (unrelated) and 4 (synonymous). Both datasets contain nouns, verbs and adjectives.

5.2 Procedure

Starting from a DM model, we matricize it into a word by link-word space ($W \times LW$) and compute similarities between words with Cosine similarity. In Exp. 1, we compute the semantic similarities of the target with each candidate and predict the candidate with the highest similarity to the target. For phrasal candidates, we compute the similarity between the target and all constituent words and take the maximum. We follow Mohammad et al. (2007) in assigning partial credit to a model when the candidates of a target are tied for maximal similarity. We evaluate the models on Exp. 2 by calculating the strength of the correlation between the model predictions and the

[3] Available from: `http://goo.gl/PN42E`

[4] Available from: `http://goo.gl/3Dflf1`

human relatedness judgments. We use Pearson's correlation coefficient since it is the de facto evaluation measure in relevant earlier work.[5]

On both tasks, we compare the models in two conditions. In the first condition ("All"), models are forced to make predictions for all items in the dataset even if they have no information about the item. In the second condition ("Covered"), models are allowed to abstain in the case of zero similarities. For Exp. 1, we report the accuracy (the number of correctly recognized synonyms divided by the number of attempted problems) and coverage (the ratio of items attempted; always 1 for the "All" condition). Items are considered covered if at least one candidate has a non-zero similarity to the target. In Exp. 2, we measure the correlation between the semantic similarities and human judgments for word pairs. Coverage is calculated as the percentage of items with similarity greater 0.

Differences between models are tested for significance using bootstrap resampling (Efron and Tibshirani, 1993), always in the "All" condition.

5.3 Models

We consider three types of DM models (monolingual, crosslingual and multilingual), bag-of-words models and a set of models proposed in the literature.

Monolingual model. We use DM.DE (Padó and Utt, 2012), constructed from a 900M-token web corpus, SDEWAC, parsed with MATE (Bohnet, 2010).[6] As discussed in Section 3.5, we consider all links (AllL) for the monolingual model.

Crosslingual models. The starting point for the crosslingual models is Baroni and Lenci (2010)'s English TypeDM model extracted from approximately 3B tokens of Wikipedia and web corpus text parsed with MaltParser (Nivre, 2006).[7] DM.XL *naive* implements Eq. (1), and DM.XL *filter* implements Eq. (3). As our translation lexicon, we use the community-built English–German `dict.cc` online dictionary.[8]

[5]We note that since the data are not normally distributed, a non-parametric correlation coefficient would be more appropriate. While we omitted them due to space limitations in this paper, we will provide Spearman ρ results for all models online at `http://goo.gl/uxuffp`.

[6]Available from `http://goo.gl/H6gViT`.

[7]Available from `http://goo.gl/63ajCI`.

[8]Available from `http://goo.gl/re44Hg`.

	Adj	Noun	Verbs	Total
English	37K	78K	8K	123K
German	35L	99K	9K	143K
Translation pairs	77K	172K	28K	277K

Table 3: Size of the `dict.cc` dictionary

Class	Model	Nodes	Edges
monolingual	DM.DE (DE)	3.5M	78M
	TYPEDM (EN)	31K	131M
crosslingual & multilingual (DE)	DM.XL naive	63K	5B
	DM.XL filter	63K	1.7B

Table 4: Sizes of various DM resources

The statistics of the dictionary in Table 3 show that it is quite large and covers many adjectives and nouns, but relatively few verbs. We had to exclude much verbal data due to ill-structured entries or phrasal entries. Following Section 3.5, we only consider selectional preference links (SPrfL) for the crosslingual model.

Multilingual models. We consider the two models described in Section 4, namely DM.MULTI Backoff and DM.MULTI MaxSim, each combining DM.DE (AllL) with DM.XL *filter* (SPrfL).

Bag-of-words models. We build a standard BOW model from the same German corpus SDEWAC used for DM.DE. We assume a window of 10 context words to the left and right. We use the top 10K most frequent content words (nouns, adjectives, verbs and adverbs) as dimensions. Our second BOW model (BOW PCA_{500}) was reduced to 500 dimensions by applying principle component analysis, a technique generally used to increase robustness to parameter choice and to combat sparsity.[9]

Models from the literature. We compare our models against the state of the art, represented by the respective best models from two previous studies (Zesch et al., 2007; Mohammad et al., 2007). They comprise monolingual ontology-based models that use GermaNet, (German) Wikipedia, or both (Lin_{GN},

[9]We also built models using smaller context windows and Latent Semantic Analysis (LSA, Landauer, 1997), both with 500 dimensions and with an automatically optimized number of dimensions (Wild et al., 2008). Since these spaces did not consistently yield better results than the reported models using PCA, we do not report the results in detail.

Model	All Acc	Covered Acc	Covered Cov
Baselines and word-based DSMs			
1 Random	.25	.25	**1**
2 Frequency	.31	.31	**1**
3 BOW	.46	.46	.98
4 BOW PCA_{500}	**.55**	**.55**	.98
Syntax-based DSMs			
5 DM.DE (AllL)	.48	.53	.84
6 DM.XL EN→DE naive (SPrfL)	.47	**.63**	.58
7 DM.XL EN→DE filter (SPrfL)	.46	.61	.58
8 DM.MULTI Backoff(7,5)	.54	.58	**.89**
9 DM.MULTI MaxSim(7,5)	**.55**	.59	**.89**
Models from the literature			
10 Lin_{dist} [MGHZ07]	NA	.52	**.45**
11 HPG [MGHZ07]	NA	**.77**	.22
12 JC [MGHZ07]	NA	.44	.36

Table 5: Exp. 1: Accuracy and Coverage for synonym choice on the Reader's Digest Word Choice dataset. MGHZ07: Mohammad et al. (2007). Best results for each model class in bold.

Model	All Corr	Covered Corr	Covered Cov
Baselines and word-based DSMs			
1 Frequency	.13	.13	**1**
2 BOW	.20	.21	.97
3 BOW PCA_{500}	**.34**	**.37**	.97
Syntax-based DSMs			
4 DM.DE (AllL) [PU12]	.38	.43	.60
5 DM.XL EN→DE naive (SPrfL)	.28	.45	.49
6 DM.XL EN→DE filter (SPrfL)	.33	**.49**	.49
7 DM.MULTI Backoff(6,4)	.40	.45	**.69**
8 DM.MULTI MaxSim(6,4)	**.42**	.47	**.69**
Models from the literature			
9 Lin_{GN} [MGHZ07]	NA	.50	.26
10 Lin_{dist} [MGHZ07]	NA	**.51**	.26
11 JC_{GN}+PL_{WP} [ZGM07]	NA	**.59**	**.33**

Table 6: Exp. 2: Coverage and correlation (Pearson's r) for predicting word similarity on the Gur350 dataset. MGHZ07: Mohammad et al. (2007)[8], ZGM07: Zesch et al. (2007)[9], PU12: Padó and Utt (2012). Best results for each model class in bold.

HPG, JC, PL); and crosslingual distributional models that represent the meaning of German lemmas in terms English thesaurus categories (Lin_{dist}).

DM model statistics. Table 4 shows the sizes of the various DMs. The German and English monolingual DMs are markedly different: the English DM is much more compact, covering only 30K lemmas while the German DM covers 3.5M lemmas, and at the same time much denser. This discrepancy is due to the larger English corpus and the inclusion of very low-frequency items in DM.DE. The crosslingual models created from TYPEDM cover 63K lemmas in German, about twice the English coverage but still almost two orders of magnitude below the monolingual DM.DE. They become very large: naive translation increases the number of edges by a factor of 30, and filtered translation still by a factor of 13. This means filtering does reduce the size of the resulting DM, but there is still considerable overgeneration.

6 Experimental Evaluation on German

The experimental results for the two experiments are shown in Tables 5 and 6, structured by model type. We observe similar patterns for the two experiments.

Baselines and word-based DSMs. In both cases, uninformed baselines (random and frequency) perform badly. (In Exp. 1, the frequency baseline predicts the most frequent item as synonym; in Exp. 2, it predicts $min(f(w_1), f(w_2))$.) In contrast, word-based DSMs perform quite well, particularly the dimensionality-reduced model (BOW PCA).

Syntax-based DSM. We see a consistent quality versus coverage tradeoff among the different classes of syntax-based DSMs. The monolingual DM.DE model is significantly outperformed by the BOW model on Exp. 1 ($p<0.01$), but numerically outperforms it on Exp. 2 (difference not significant).

In both tasks, the crosslingual DM.XL models outperform both DM.DE and BOW PCA in terms of quality: They achieve the numerically highest accuracy (and correlation, respectively) among all syntax-based models. This high quality comes at a low coverage, matching our intuitions about the profile of the

[8]Mohammad et al. (2007) do not provide coverage numbers in their paper. We appreciate the support of Torsten Zesch and Saif Mohammad in recovering the necessary information.

[9]Zesch et al. (2007) report results for the subset of Gur350 in the intersection of GermaNet and Wikipedia. Thus, their models may have higher coverage on the complete Gur350, but to our knowledge these numbers have not been published.

crosslingual model. Filtering leads to a significant improvement in Exp. 2 ($p<0.05$) but not in Exp. 1.

The multilingual models (DM.MULTI) perform even better. They nearly retain the quality of the crosslingual models (accuracy of .59 vs. .63 for Exp. 1, correlation of .47 vs. .49 for Exp. 2) but attain higher coverage (89% in Exp. 1 and 69% in Exp. 2) Notably, the coverage is even higher than that of the DM.DE models, attesting to the complementarity of mono- and crosslingual information.

The differences among the DM.MULTI models are small, but MaxSim does a little better and performs best overall. In Exp. 1, it does significantly better in the all-items evaluation than all other syntax-based models ($p<0.01$). The differences in Exp. 2 are only significant at $p<0.05$ for the model pair 6–8; we attribute this to the smaller size of the dataset.

In sum, we can construct crosslingual DMs without any use of target language corpora that mirror or even exceed the performance of monolingual DMs. If monolingual data is available, the combination of corpus evidence provides a substantial advantage over both monolingual and crosslingual models, even for German, a language with large, relatively reliably parsed corpora. Users can choose among different models with different coverage/quality tradeoffs.

Comparison to models from the literature. Models from the literature are shown at the bottom of the two tables. They generally obtain the highest accuracy (or correlation, respectively), but only cover a relatively small part of the datasets. In particular, the models with a quality higher than the DM variants (11 in Exp. 1 and 10 and 11 in Exp. 2) exhibit a coverage of less than half than that of the DM.MULTI models. This appears to show the usual trade-off between hand-constructed knowledge and automatically acquired knowledge (Gildea and Jurafsky, 2002). However, we can similarly bias our DMs towards accuracy with the aid of a simple frequency filter that only permits predictions for items where all involved lemmas occur more frequently in the German corpus than some threshold. Setting these thresholds to match the coverage figures of the best ontology-based models, DM.MULTI MaxSim almost reaches the ontology-based results: On Exp. 1, for a coverage of .22 we obtain an accuracy of .68 (ontology-based model: .77), and on Exp. 2, we obtain a correlation of .60 (ontology-based model: .59) at a coverage of .33.[10] Thus, our DM models approximate the quality of ontology-based models without using any handcrafted resources.

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

Table 7: Words of foreign origin better represented by the multilingual model

Differences between Exp. 1 and 2. The two main differences between the experiments are (a) the performance of DM relative to the BOW baseline, and (b) the impact of backtranslation filtering. In Exp. 1, the BOW performs as well as DM.MULTI, and the unfiltered DM.XL has a slight edge (2% accuracy) over the filtered one. In contrast, in Exp. 2 filtering leads to a major improvement and DM.MULTI does substantially better than BOW PCA. Our analyses attribute this difference to the nature of the two tasks (cf. Section 5). Exp. 1 requires the recognition of synonyms. Here, the main determinant of success is whether the actual synonym receives the highest similarity or not, irrespective of the margin to the competing candidates. This margin does increase from 0.09 in the naive to 0.11 in the filtered DM.XL, but the overall number of correct predictions remains almost unchanged. In contrast, Exp. 2 covers the whole range from highly similar to unrelated word pairs, and the correlation evaluation is sensitive to the relative size of similarities produced by the DM across many word pairs. The improvement we see indicates that filtering improves the overall scaling of the similarities, but this effect is masked by the decision criterion in Exp. 1.

Qualitative analysis. Comparing DM.DE with DM.MULTI, the question arises: can we further characterize the benefits that the inclusion of crosslingual corpus evidence confers to monolingual models? We first inspected Exp. 1 for synonyms that were correctly recognized by DM.MULTI MaxSim but not DM.DE, and found a large number of words of foreign origin (see Table 7). These words tend to be rare in the German corpus in the form of technical, slang,

[10] We cannot provide significance tests since we do not have item-wise predictions for the models from the literature.

or elevated register terms. Due to their low level of ambiguity as well as the fact that their English translations are often more frequent, the crosslingual model represents them more sensibly.

We then inspected Exp. 2 in a similar way but found it more difficult to identify salient improved classes since the improvement is mostly in terms of coverage. The data set for Exp. 2 includes proper nouns, such as *Berlin/Berlin-Kreuzberg, Benedetto/Benedikt*, which are unlikely to be covered by a translation lexicon. It also contains items that encode world knowledge such as *Ratzinger/Papst (pope)* which has a better chance of being covered by target language corpora. This pair is not covered by the DM.XL models, but the monolingual models (DM.DE, BOW, and BOW PCA) assign it the similarities .23, .66, and .89, respectively.

7 Experimental Evaluation on Croatian

Our third experiment considers a language that is more different from English than German, namely Croatian, a Slavic language. Available resources for Croatian are more limited than for English or German. Since syntactic analysis used to be a bottleneck, the first syntax-based DSM for Croatian, DM.HR, became available only last year (Šnajder et al., 2013). As for evaluation datasets, there are no human similarity judgments, but there is a synonym choice dataset (CroSyn – see Karan et al. (2012) for details).

Thus, our Croatian evaluation is a synonym choice task parallel to Exp. 1 for German. We take DM.HR as the monolingual model which was built from a dependency-parsed Croatian web corpus of 1.2B tokens. We construct a crosslingual model by starting from Baroni and Lenci's English TypeDM and using Taktika Nova's freely available English–Croatian dictionary[11] with 105K translation pairs. After removing entries with more than one word per language, we were left with 95K pairs, considerably fewer than for English–German. We apply the methods from Section 3 for edge translation and filtering. The resulting filtered Croatian DM.XL has 47K nodes and 315M edges, about one order of magnitude smaller than the German crosslingual resource. Finally, we combined DM.HR with the crosslingual DM (as in Section 4) to obtain multilingual Croatian DMs.

[11] Available from `http://goo.gl/xHUjJH`

Model	All Acc	Covered Acc	Covered Cov
Word-based DSMs			
1 BOW-LSA [SPA13]	.66	.66	**1**
Syntax-based DSMs			
2 DM.HR (AllL)	.65	.65	**.99**
3 DM.XL EN→HR naive (SPrfL)	.43	.50	.71
4 DM.XL EN→HR filter (SPrfL)	.58	**.71**	.71
5 DM.MULTI Backoff(4,2)	.69	.69	**.99**
6 DM.MULTI MaxSim(4,2)	**.70**	.70	**.99**

Table 8: Experiment 3: Accuracy and Coverage for synonym choice on the CroSyn dataset. SPA13: Šnajder et al. (2013). In boldface: best results.

Table 8 shows the results which correspond closely to those for Exp. 1. A dimensionality-reduced BOW space performs competitively with the monolingual DM.HR (Šnajder et al., 2013). The crosslingual DM is again able to improve accuracy over DM.HR (by 6%) but drops in coverage. Again, the multilingual models perform best: DM.MULTI MaxSim loses only 1% accuracy compared to the crosslingual model but achieves almost perfect coverage. The differences to DM.HR and DM.XL are both significant ($p<0.01$).[12]

The two major differences to the German synonym choice task (Exp. 1) are that (a) filtering plays an essential role for Croatian (increase in accuracy by 15%) and (b) DM.MULTI clearly outperforms the BOW model. We attribute the difference to the semi-automatic construction of the Croatian dataset from machine-readable dictionaries. Overall, the results for Croatian are encouraging. They demonstrate that languages where parsing technology is still developing can in particular profit from cross- and multilingual methods. This is true even for relatively small translation dictionaries, matching previous results from the literature (Peirsman and Padó, 2011).

8 Related Work

Given the resource gradient between English and other languages, the crosslingual induction of linguistic information has been an active topic of research.

Many studies use parallel corpora. Annotation projection (Yarowsky and Ngai, 2001) transfers source language annotation directly onto target language

[12] We cannot provide significances for the BOW results because we again do not have per-item predictions.

sentence. It has been applied to various linguistic levels such as POS tagging and syntax (Hi and Hwa, 2005; Hwa et al., 2005, among others). Other studies use parallel data as indirect supervision for monolingual tasks. Diab and Resnik (2002) use translations as word sense labels; van der Plas and Tiedemann (2006) exploit multilingual distributional semantics for robust synonymy extraction. Naseem et al. (2009) learn unsupervised POS taggers on multilingual parallel data, exploiting the differences between languages as soft constraints. Titov and Klementiev (2012) and Kozhevnikov and Titov (2013) induce shallow semantic parsers from parallel data. Klementiev et al. (2012) approach document classification with multitask learning, inducing a multilingual DSM.

Since parallel corpora are not available in large quantities, other studies use *comparable corpora* which can provide additional features from the other language. For example, Merlo et al. (2002) improve English verb classification with new features derived from Chinese translations. De Smet and Moens (2009) learn multilingual topic models for news aggregation. Peirsman and Padó (2011) use comparable corpora to transfer selectional preferences and sentiment labels. Wikipedia can be seen as a particularly rich type of comparable corpus with additional link structure. It has been used to compute semantic relatedness (Navigli and Ponzetto, 2012; Navigli and Ponzetto, 2010) and to compute conceptual document representations for crosslingual information retrieval (Potthast et al., 2008; Cimiano et al., 2009).

Our work, does not require parallel or comparable corpora. We note, however, that translation lexicons such as the ones we use can be extracted from comparable corpora (Rapp, 1999; Vulić and Moens, 2012, and many others), though few papers are concerned with the translation at the level of semantic relations, as we are. Similar in this respect is Fung and Chen (2004), who translate FrameNet (Baker et al., 1998) into Chinese with a bilingual ontology. They use a relation-based pruning scheme that is somewhat comparable to our backtranslation filtering.

To our knowledge, the most similar work to ours is (Mohammad et al., 2007), which also considers DSMs, albeit a different variety, namely *concept-based* DSMs where targets are characterized in terms of their distribution over categories of Roget's thesaurus. Like our work, their study creates crosslingual DSMs for German using a translation lexicon. It follows a different strategy, however: it collects co-occurrence counts from a German corpus and translates the context dimensions into the English Roget categories. Therefore, it crucially requires a large target language corpus, which our crosslingual methods (Section 3) avoid. Its use of a target language corpus resembles our multilingual methods (Section 4), but unlike them, does not combine corpus evidence from both languages. In sum, we believe that our methods are more adaptable to different scenarios, being able to use whatever data is available in either language.

9 Conclusion

The appeal of syntax-based distributional spaces lies in their promise of flexible and linguistically more appropriate models for many phenomena in lexical semantics. A major obstacle to their adoption for novel languages has been the significantly higher requirements on resources compared to word spaces.

In this paper, we have demonstrated that this obstacle can be overcome by transferring English distributional information along the resource gradient into target languages such as German and Croatian. The simplest models, which are based solely on the English Distributional Memory (DM) resource and a translation lexicon, already beat monolingual DMs in quality. These crosslingual models suffer from lower coverage but can be combined with the monolingual DM yielding a multilingual DM that maintains competitive accuracy while achieving significantly higher coverage than either individual model. The outcomes of our experiments are mostly stable across the languages and tasks presented, which leads us to assume the methodology successfully generalizes.[13]

Directions for future research include (a), more stringent filtering of spurious edges in DM.XL models to make the graph topology more similar to monolingual models and enable graph merging to obtain unified multilingual models; (b), the extension of our approach to more than two languages; (c), dimensionality reduction for tensor-based DSMs both for efficiency reasons and to improve performance.

[13]The German DMs are publicly available from `http://goo.gl/uxuffp`.

Acknowledgments

We gratefully acknowledge partial funding of our research by the DFG (SFB 732, Project D6) and the EC (Project EXCITEMENT, FP7 ICT-287923).

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, pages 86–90, Montréal, QC.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):1–49.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2008. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing, China.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164, New York, NY.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Center for Research in Computing Technology, Harvard University.

Philipp Cimiano, Antje Schultz, Sergej Sizov, Philipp Sorg, and Steffen Staab. 2009. Explicit vs. latent concept models for cross-language information retrieval. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1513–1518, Pasadena, CA.

Wim De Smet and Marie-Francine Moens. 2009. Cross-language linking of news stories on the web using interlingual topic modelling. In *Proceedings of the CIKM Workshop on Social Web Search and Mining*, pages 57–64, Hong Kong.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 255–262, Philadelphia, PA.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York, NY.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics*, 36(4):723–763.

Pascale Fung and Benfeng Chen. 2004. BiFrameNet: Bilingual Frame Semantics Resources Construction by crosslingual Induction. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 931–935, Geneva, Switzerland.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston/Norwell, MA.

Zelig S. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Chenhai Hi and Rebecca Hwa. 2005. A Backoff Model for Bootstrapping Resources for Non-English Languages. In *Proceedings of the joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 851–858, Vancouver, BC.

Rebecca Hwa, Philipp Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Journal of Natural Language Engineering*, 11(3):311–325.

Eric Joanis, Suzanne Stevenson, and David James. 2006. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(03):337–367.

Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Distributional semantics approach to detecting synonyms in Croatian language. In *Proceedings of the Eighth Language Technologies Conference*, Ljubljana, Slovenia.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of the International Conference on Computational Linguistics*, pages 1459–1474, Mumbai, India.

Mikhail Kozhevnikov and Ivan Titov. 2013. Crosslingual transfer of semantic role models. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*, pages 1190–1200, Sofia, Bulgaria.

Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Dekang Lin. 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*, pages 768–774, Montreal, QC.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan 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*, pages 523–530, Vancouver, BC.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97, Sofia, Bulgaria.

Paola Merlo, Suzanne Stevenson, Vivian Tsang, and Gianluca Allaria. 2002. A multilingual paradigm for automatic verb classification. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 207–214, Philadelphia, PA.

Rada Mihalcea and Dragomir Radev. 2011. *Graph-based Natural Language Processing and Information Retrieval*. Cambridge University Press, Cambridge, UK.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Saif Mohammad, Iryna Gurevych, Graeme Hirst, and Torsten 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*, pages 571–580, Prague, Czech Republic.

Tahira Naseem, Benjamin Snyder, Jacob Eisenstein, and Regina Barzilay. 2009. Multilingual Part-of-Speech Tagging : Two Unsupervised Approaches. *Journal of Artificial Intelligence Research*, 36:1–45.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelrelate! a joint multilingual approach to computing semantic relatedness. In *Proceedings of the 26th Conference on Artificial Intelligence*, pages 108–114, Toronto, ON.

Joakim Nivre. 2006. *Inductive Dependency Parsing*. Springer, Dordrecht, Netherlands.

Sebastian Padó and Jason Utt. 2012. A distributional memory for German. In *Proceedings of the KONVENS 2012 workshop on recent developments and applications of lexical-semantic resources*, pages 462–470, Vienna, Austria.

Yves Peirsman and Sebastian Padó. 2011. Semantic relations in bilingual lexicons. *ACM Transactions in Speech and Language Processing*, 8(2):3:1–3:21.

Lonneke van der Plas and Jörg Tiedemann. 2006. Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, pages 866–873, Sydney, Australia.

Martin Potthast, Benno Stein, and Maik Anderka. 2008. A wikipedia-based multilingual retrieval model. In *Proceedings of the European Conference on Information Retrieval*, pages 522–530, Glasgow, Scotland.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526, College Park, MD.

Sabine Schulte im Walde. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, 32(2):159–194.

Hinrich Schütze. 1992. Dimensions of meaning. In *Proceedings of Supercomputing '92*, pages 787–796, Minneapolis, MN.

Stephen Soderland, Oren Etzioni, Daniel S Weld, Michael Skinner, Jeff Bilmes, et al. 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*, pages 262–270, Suntec, Singapore.

Harold Somers. 2005. Round-trip translation: What is it good for? In *Proceedings of the Australasian Language Technology Workshop*, pages 127–133, Sydney, Australia.

Jan Šnajder, Sebastian Padó, and Željko Agić. 2013. Building and evaluating a distributional memory for Croatian. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 784–789, Sofia, Bulgaria.

Ivan Titov and Alexandre Klementiev. 2012. Crosslingual induction of semantic roles. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 647–656, Jeju Island, South Korea.

Peter Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Ivan Vulić and Marie-Francine Moens. 2012. Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 449–459, Avignon, France.

DeWitt Wallace and Lila Acheson Wallace. 2005. *Reader's Digest, das Beste für Deutschland*. Verlag Das Beste, Stuttgart, Germany.

Fridolin Wild, Christina Stahl, Gerald Stermsek, and Gustaf Neumann. 2008. Parameters driving effectiveness of automated essay scoring with LSA. In *Proceedings of the 9th Computer-Aided Assessment Conference*, pages 485–494, Loughborough, UK.

David Yarowsky and Grace 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*, pages 200–207, Pittsburgh, PA.

Torsten Zesch, Iryna Gurevych, and Max 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*, pages 205–208, Rochester, NY.