Multilingual generative models for selectional preference learning

Adrian Scoică
Girton College

UNIVERSITY OF
CAMBRIDGE

A dissertation submitted to the University of Cambridge in partial fulfilment of the requirements for the degree of Master of Philosophy in Advanced Computer Science (Option B)

University of Cambridge
Computer Laboratory
William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom

Email: as2270@cl.cam.ac.uk

June 13, 2013

Declaration

I Adrian Scoică of Girton College, being a candidate for the M.Phil in Advanced Computer Science, hereby declare that this report and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose.

Total word count: 14,502

Signed:

Date:

Acknowledgements

I would like to thank Dr. Diarmuid Ó Séaghdha, for his constant guidance and valuable feedback throughout the year. This project would not have been possible without him. I would also like to give thanks to Dr. Rada Mihalcea, Dr. Sebastian Padó, and Radu Simionescu for our insightful e-mail exchanges.

Last but not least, I would like to thank my friends for being there for me.

Abstract

This dissertation investigates the effectiveness of cross-lingual transfer of verb-argument selectional preferences from English to resource-poor languages in which little or no dependency-parsed data is available for training monolingual plausibility estimation models.

In the first stage of the research, four monolingual models of selectional preference, including two generative topic models based on Latent Dirichlet Allocation, were trained on dependency-parsed corpora in English, German, Spanish, and Romanian, and evaluated by correlating their predictions with gold standards of human judgements. In the second stage, the cross-lingual plausibility transfer using bilingual vector spaces was tested from English into the other three languages.

The results show that LDA models outperform the other monolingual models, and that cross-lingual models trainable with POS-tagged data can be better than monolingual methods trained on small corpora. Furthermore, augmentations were proposed to help topic models deal with unknown words, and to build better bilingual vector spaces. Lastly, the research identified key limitations of cross-lingual transfer and possible directions for future research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem description and motivation

Selectional preference is a linguistic phenomenon which describes the *affinities* or *restrictions* that words may have on the type of arguments (such as *objects* or *subjects* in the case of verbs) they can take in a given language model. The violation of selectional preferences can lead to sentences which, while grammatical, are considered unacceptable. The following famous example composed by Noam Chomsky in (Chomsky, 1957) illustrates this problem:

"Colourless green ideas sleep furiously."

This sentence sounds implausible because selectional preferences are violated for all grammatical relations: *'idea'* denotes an abstract concept; it is neither among the types of entities which can typically *'sleep'*, nor can it take modifiers specifying physical properties such as colour. Furthermore, *'sleep'* denotes a static process, while modifier *'furiously'* normally describes dynamic, ongoing actions, and *'colourless'* and *'green'* are, from a logical standpoint, mutually exclusive properties.

The subject of selectional preferences has been widely studied in the literature, and the proposed approaches to modelling them fall into two main categories: class-based approaches, and non-class approaches. The former try to map arguments onto a predefined class taxonomy, which is an expensive resource that needs to be designed separately, while the latter avoid this need by either computing similarity measures without relying on an ontology, or by automatically inducing the set of classes. Of the two, non-class approaches are appealing because they can be automatically trained on dependency-parsed corpora to produce predictions which correlate well with estimates made by human judges.

However, since previous work on selectional preference induction has focused mainly on languages where large parsed corpora can be compiled, such as English, these monolingual methods cannot be applied to languages for which such resources are not available. This could happen either because no parser is available or accurate enough for a given language, or because no efforts have been made to compile a large text corpus in that language.

Based on the observation that selectional preferences are a consequence of world facts, and thus should preserve across languages, this project investigates how well the translation-based transfer of selectional preferences from resource-rich to resource-poor languages works. Furthermore, the project aims to determine the threshold for the amount of dependency-parsed training data beyond which cross-language plausibility transfer overtakes monolingual methods, along with the robustness of those methods.

A representative set of results is given in figure 1.1. The graph shows that the automated cross-lingual transfer from English using bilingual vector models (plotted in brown) outperforms the monolingual methods (plotted in green, red, blue and purple), confirming that cross-lingual selectional preference transfer can be applied to resource-poor languages where monolingual methods cannot be trained.

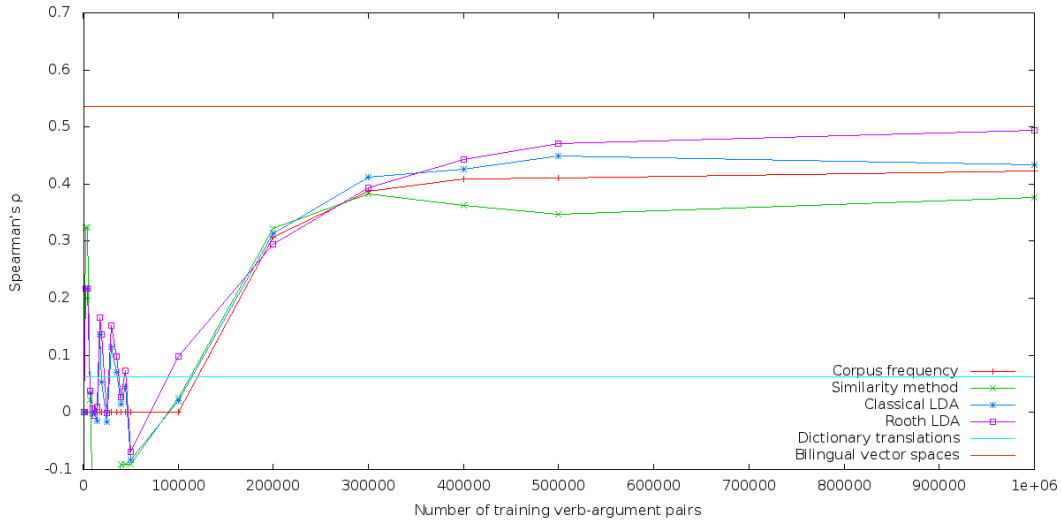


Figure 1.1: Results sample, showcasing the performance of the studied four monolingual and two cross-lingual models on German verb-subject selectional preferences.

1.2 Dissertation plan

Chapter 2 describes the main previous approaches to computing and evaluating selectional preference plausibilities. Chapter 3 describes the monolingual and cross-lingual algorithms I implemented and tested, along with the methods used for estimating the performance baselines. Chapter 4 contains a detailed description of the data resources used for training the algorithms.

Chapter 5 is organized in four parts. In sections 5.1 and 5.2, I explain the testing methodology and outline the test datasets used. In sections 5.3 and 5.4, I compare performances with two previous studies, discuss observations, and describe my suggested improvements. In section 5.5, I determine the threshold size of the dependency-parsed training corpus for obtaining better performances with cross-lingual plausibility transfer than with monolingual models, and in section 5.6 I discuss the main observations from the error analysis stage.

Finally, Chapter 6 gives a summary of the findings, and indicates possible directions for future research.

Chapter 2

Background and related work

2.1 Selectional preferences

2.1.1 Applications

Selectional preferences have a large number of applications in Natural Language Processing. The most widely cited ones are semantic role labeling (Gildea and Jurafsky, 2002), disambiguation tasks and resolution tasks. Two important examples of using selectional preferences for disambiguation tasks are Word Sense Disambiguation (McCarthy and Carroll, 2003) and syntactic disambiguation (Hindle and Rooth, 1993), while resolution tasks include pronoun resolution (Bergsma et al., 2008) and syntactic, word sense and reference ambiguity resolution (Clark and Weir, 2002).

2.1.2 Main approaches

Statistical approaches to inducing selectional preferences began in the Natural Language Processing literature in the 1990s, with publications such as (Grishman and Sterling, 1993) and (Resnik, 1993). While the earliest attempts were class-based, the main disadvantage of the class-based approaches

is that they require an ontology, which is an expensive resource that needs to be designed separately. To avoid this bottleneck, non-class approaches of modelling selectional preference have recently been proposed.

One solution described in (Bergsma et al., 2008) is the discriminative approach of training a Support Vector Machine classifier to distinguish between positive examples collected from the observed predicate-argument pairs in the data and artifically constructed negative examples.

Another type of non-class methods are the similarity-based approaches, the principles of which were described in (Dagan et al., 1999) and in (Grishman and Sterling, 1993) when describing smoothing techniques for addressing the data sparsity problem of word co-occurrence probability calculation. The selectional preference of a relation rel for a possible heardword arg is then computed as a weighted sum of the similarities between arg and the other words seen in relation rel, as described in equation 2.1.

$$P_{vb,rel}(arg) = \sum_{arg' \in Seen(vb,rel)} sim(arg, arg') wt_{vb,rel}(arg') \qquad (2.1)$$

(Erk, 2007) describes an implementation of one such a model for the automatic induction of selectional preferences using corpus-based similarity metrics. The implementation uses two corpora: a primary one for extracting predicate-argument frequency data, and a secondary one for computing a semantic similarity metric. The method is shown to have lower error rates than the WordNet-based model of (Resnik, 1993) and the later model of (Rooth et al., 1999), but to suffer from lower coverage.

The most recent approaches to modelling selectional preferences are generative probabilistic models, which model verb arguments as having been randomly generated by latent variables. One method of automating selectional preference learning from unlabelled data using generative models is described in (Rooth et al., 1999). This approach models the semantic classes of arguments as hidden variables which are derived from distributional data obtained from parsing an unannotated corpus. The verb $vb \in V$ and its argu-

ment $arg \in N$ are considered to be conditioned on a hidden class $class \in C$, and the selectional preference is computed as a smoothed probability given by equation 2.2. While Rooth proposes training the model with EM-based clustering, another training alternative is to use Gibbs sampling.

$$p(vb, arg) = \sum_{class \in C} p(class, vb, arg) = \sum_{class \in C} p(class)p(vb|class)p(arg|class) \tag{2.2}$$

Probabilistic topic models, on the other hand, are statistical latent variable models which have been *traditionally* used to model document-word co-occurrences, but which have recently been applied to a variety of tasks in Natural Language Processing, including the modelling of selectional preferences. Based on the intuition that predicates select arguments from a number of classes that are shared across all predicates, the power of topic models lies in the ability to automatically induce within a solid mathematical framework both the semantic classes and their distributions over predicates.

2.1.3 Evaluation methods

The literature contains descriptions of multiple frameworks developed for evaluating the performance of selectional preference plausibility predictions.

Some of the earlier reported methods measure the correlation between the predictions output by the algorithms and a gold standard of human plausibility judgements. The methodology employed for selecting the verb-argument pairs on which to elicit human judgements, along with the resulting gold standards, were presented in (Keller and Lapata, 2003) and (Brockmann and Lapata, 2003) for English and German, respectively. This evaluation method was later employed for most plausibility estimation approaches, such as similarity-based models in (Erk, 2007), topic models in (Ó Séaghdha, 2010), and cross-lingual transfer in (Peirsman and Padó, 2010).

Alternative, but less frequently used evaluation techniques include the use

of pseudo-words, and indirectly measuring the performance through the performance of another task which uses the estimates as inputs. The former, described in (Chambers and Jurafsky, 2010), was originally developed for word sense evaluation, and works by proposing a replacement for the argument in a verb-argument pair, and asking the system to identify the original. The latter was used in (Bergsma et al., 2008), and relies on the performance of pronoun resolution to measure the quality of the selectional preference predictions.

2.2 Topic models

2.2.1 Overview

The most widely cited topic models in Natural Language Processing are Latent Semantic Indexing (LSI), Probabilistic Latent Semantinc Indexing (pLSI) and Latent Dirichlet Allocation (LDA). These methods are used in Information Retrieval to produce descriptions of the documents in a large collection, which would enable their efficient processing.

The seminal work on Latent Dirichlet Allocation was formally described in (Blei et al., 2003), which describes the assumptions and mathematical background of the model, parameter inference and smoothing techniques, as well as potential applications. The subject was revisited in (Steyvers and Griffiths, 2007) from a more applied perspective, as the authors propose the Gibbs sampling algorithm as a tractable method of inferring the parameters of the model. Probabilistic topic models are further discussed in (Blei, 2012), which gives the details of a concrete application to document classification. The paper also proves the usefulness of LDA in producing semantically-coherent, human-interpretable topics associated to the automatically induced classes.

The main strength of topic models over other document classification algorithms is that documents are viewed as a mixture of topics, who are themselves probability distributions over words. One advantage of this approach

is that it allows topic models to generalize well to documents unseen in the training corpus. Another key advantage is that words can belong to multiple topics, which allows the model to deal with polysemy.

2.2.2 Application to selectional preferences

Topic models address an important intuitive requirement of automatic selectional preference learning: they are naturally adapted to handling sparse data, which is the most important challenge in training selectional preference models.

(Ritter et al., 2010) proposed three variations of basic LDA (IndependentLDA, JointLDA and LinkLDA) for unsupervised selectional preference learning by varying the conditional link between the predicate and the argument, and achieved state of the art performance with a model in which the predicate and the argument are generated by distinct latent variables sampled from the same document topic distribution. These results confirmed that topic models are a viable solution to modelling selectional preferences. At the same time, a comparison of three different approaches based on topic models described in (Ó Séaghdha, 2010) confirmed that they can be adapted to selectional preference learning and that for infrequent predicate-argument combinations, they can outperform classical methods based on corpus counting.

2.3 Multilingual knowledge transfer

Most of the state of the art on selectional preferences has been carried out on languages for which large annotated corpora exist, such as English. Although, intuitively, the phenomenon of selectional preference is to a great degree dependent on the state of affairs in the world, rather than being purely a feature of the language, there is currently no solution to this problem for resource-poor languages.

Previous research indicates that multilingual knowledge transfer is possible

for various tasks in Natural Language Processing. The 5th Workshop on Semantic Evaluation[1] defined frameworks for evaluating the performance of solutions to, among others, a cross-lingual lexical substitution task (Mihalcea et al., 2010), and a cross-lingual word sense disambiguation task (Lefever and Hoste, 2010).

A transfer of semantic knowledge between pairs of languages has previously been studied in (Peirsman and Padó, 2008). A later study (Peirsman and Padó, 2010) by the same authors focused on the automated cross-language learning of selectional preferences from unrelated corpora between pairs of languages. The algorithm they propose relies on building a bilingual vector space to translate predicate-argument pairs between two languages, and then relies on the vector space to approximate the plausibility of predicate-argument combinations in the resource-poor language by their translations. The process is described by equation 2.3, where Pl stands for the plausibility of a predicate p taking the head-word h for an argument a, the superscripts s and t indicate the source and target languages, respectively, and tr is the translation function.

$$Pl^s(p, a, h) = Pl^t(tr(p), a, tr(h)) \tag{2.3}$$

The authors also specify a smoothing technique within unilingual methods, as described by equation 2.4. The variable notations used are the same as in equation 2.3, while function $Seen$ returns the set of seen heads for argument slot a of verb v, function w is a weighting function, and function sim is a similarity function.

$$P(h|a, v) = \Sigma_{h' \in Seen(v,a)} \frac{w(h')sim(h, h')}{\Sigma_{h'} w(h')} \tag{2.4}$$

A multilingual approach to computing semantic relatedness, and thus a similarity function, has been described in (Navigli and Ponzetto, 2012), while a method for computing cross-lingual semantic relatedness using Wikipedia,

[1] http://semeval2.fbk.eu/semeval2.php

developed and described in (Hassan and Mihalcea, 2009), achieved performances comparable to those of translation engines and proving the usefulness of Wikipedia for computing language alignment.

2.4 Summary

In this chapter, I showed that selectional preferences have numerous applications in NLP, that the statistical approaches to computing selectional preferences in the literature go back to the 1990s, and that the application of topic models to selectional prefrences is a state of the art solution. I also showed that while multiple evaluation techniques were described in the literature, the correlation of plausibility predictions with a gold standard of human judgements is the most frequently used one. Last but not least, I showed that previous work confirms the usefulness of Wikipedia for computing multilingual similarity functions and language alignments.

Chapter 3

Algorithms

3.1 Monolingual models

3.1.1 Corpus frequency baseline

The most straightforward way of estimating the plausibility of a verb-argument pair in the context of a given grammatical relation is to define it as the joint probability of the verb-argument combination in the language model, and then to estimate this joint probability on the basis of corpus counts. Ideally, with access to an unlimited amount of dependency-parsed corpus data, this method of estimating probabilities would enable one to build the best language selectional preference model achievable. In practice, however, the performance of this model is directly dependent on the amount of parsed corpus data available, and since parsed corpus data is an expensive resource, the model is not generally applicable to resource-poor languages.

Despite this shortfall, I had two reasons for using the corpus frequency model as a monolingual baseline. The first reason is that it enables an objective performance comparison between languages with uneven amounts of training data for monolingual models of selectional preference: given the fact that I had different amounts of parsed data available for each of the four

languages used in my study, the simple comparison of the correlation indices between the system's output and the gold standards is not indicative of actual model performance differences. The second reason deals with the differences in the difficulty of the gold standard test datasets themselves. While the datasets described in section 5.2 were all compiled using the same principles, the datasets for some languages contain less likely triples. This makes the plausibilities more difficult to predict, and causes uneven correlation scores among languages. By comparing the performance of other models to the performance of the corpus frequency baseline model instead, I could interpret evaluation results more reliably.

Furthermore, as this baseline is particularly sensitive to decreasing the amount of training data available, the corpus frequency model allowed me to comparatively assess the robustness of the other monolingual models to data scarcity.

The model estimates the joint probability using the corpus-derived co-occurrence matrix of verbs and dependent noun heads, based on the chain rule (equation 3.1), rather than direct estimation.

$$P(vb, arg, rel) = P(arg|vb, rel) \cdot P(vb, rel) \tag{3.1}$$

The reason for using the chain rule is that breaking the joint probability in two components allows for a more flexible smoothing scheme, and a more parameterized treatment of unknown verbs or arguments. For the purpose of my baseline experiments, however, I assigned 0.00 to the probability for unknown verb-argument combinations.

3.1.2 Similarity methods

The similarity-based model for automatic induction of selectional preferences appeared as an intuitive development on the corpus frequency baseline model, and the underlying principles of the method were first described in (Erk,

2007). A similar technique was used in (Padó et al., 2007), co-authored by Erk, and later cited as a method for obtaining monolingual baselines in (Peirsman and Padó, 2010). The method is accurately explained by equation 3.2, which describes the computation of the plausibility of a given verb-argument pair as an interpolation of the conditional verb-argument probabilities of all heads seen in the training corpus in the same relation to the given verb, weighted by their similarity to the original argument for which the plausibility is being computed.

$$P(arg, vb, rel) = \sum_{arg' \in Seen(vb,rel)} P(arg', vb, rel) \cdot sim(arg, arg') \qquad (3.2)$$

The method uses two corpora: a primary corpus, and a generalization corpus. The primary corpus is used for extracting a verb-argument co-occurrence matrix for computing $P(arg', vb, rel)$, as previously described in section 3.1.1, for which reason it has to be in dependency-parsed format. The generalization corpus, on the other hand, is used for computing the corpus-based similarity metric, *sim*. Following the description given in (Peirsman and Padó, 2010) of the experiments we reproduced, I computed similarity by building a syntactic context vector space for all common nouns in the generalization corpus, and by computing similarity with vector distance metrics. For this reason, I required the generalization corpus to also be in dependency-parsed format, and I used the same corpus as the primary corpus as suggested by the original author in (Erk, 2007).

There are two main reasons why I included this method in my research. On one hand, using a similarity-based method enabled me to reproduce the monolingual baseline experiments on Spanish reported in (Peirsman and Padó, 2010), which were based on the AnCora corpus. On the other hand, however, the similarity method is appealing from the perspective of its application to resource-poor languages. While the method requires the primary corpus to be dependency-parsed, the impact of the size of the primary corpus on the quality of the predictions is less than in the case of the corpus frequency models, because more information can be extracted from the

dependency-parsed data available. This is because the formula does not rely directly on the argument being in the given relation to the verb in the training data: the fact that the syntactic context vector space includes all common nouns in the generalization corpus leads to better coverage. However, I assigned 0.00 to the probability for verb-argument combinations where either the verb or the argument is not present on its own in the training corpus.

3.1.3 Latent Dirichlet Allocation

Overview

While I used the previous monolingual algorithm described in section 3.1.1 as a baseline for evaluating the differences in the training and testing data, and the one described in section 3.1.2 for comparison with previous research, the state of the art in automatic selectional preference modelling at the moment of writing this dissertation are topic models. As previously mentioned in section 2.2, topic models lend themselves well to modelling selectional preferences, because they are naturally suited to dealing with sparse data and can be efficiently trained using Gibbs sampling.

The latent variable model that I used was Latent Dirichlet Allocation, which was initially developed for modelling document-term co-occurrences. The terms in a document are viewed as belonging to a mixture of topics. The topic mixture for the document follows a multinomial distribution drawn from a Dirichlet prior, and each topic is itself a multinomial distribution over words. According to these model specifications, a three-step generative story can be formulated for the document, allowing the distributions to be inferred through Gibbs sampling on the training set.

By analogy to the document-term co-occurrences, we can also view selectional preference as a generative process, in which the documents are replaced with predicates, the topics are called classes, and the document terms are replaced with the observed arguments that a predicate takes in the training corpus. This analogy is explained in more detail in (Ó Séaghdha, 2010). Of the LDA

varieties which can be used to model selectional preferences, I chose the standard version, which I called *Classical LDA* throughout this paper, and a version inspired from (Rooth et al., 1999), which I called *Rooth LDA*. These varieties were reported in (Ó Séaghdha, 2010) to obtain the best performances of three models tested. Following the methodology in the paper, I used the Gibbs sampling library implementations found in the MALLET machine learning toolkit to train both types of topic models and to optimize their hyperparameters. Further details about the toolkit are given in (McCallum, 2002).

The single most important disadvantage of LDA in modelling selectional preferences is its inability to deal with unknown words. While this problem led me to report performance on the maximal subset of the test dataset that achieves full coverage, I also proposed augmenting the model with a lexical substitution method based on vector space models in order to address coverage problem. The implementation details of this improvement are given in section 5.3.2.

The following two sub-sections give further details about the generative story and the plausibility calculation procedure for each model variety.

Classical LDA

In **Classical LDA**, the three-step generative story for modelling selectional preferences taken from (Ó Séaghdha, 2010) closely follows the one for document topics:

1. For each verb vb, a multinomial distribution over argument classes is drawn from a Dirichlet distribution with parameters α.

2. For each argument class, a multinomial distribution over arguments is drawn from a Dirichlet distribution with parameters β.

3. An argument for a given verb is generated by first drawing a class from the verb's class distribution, and then by drawing an argument from the argument distribution of the class.

Once we have trained the parameters of the model, we can estimate the plausibility of a verb-argument pair in the context of a given grammatical relation by using equation 3.3, where f denotes training corpus frequency counts, and N is the size of the argument vocabulary.

$$P(arg|vb, rel) = \sum_{class} P(arg|class)P(class|vb, rel) \tag{3.3}$$

$$\propto \sum_{class} \frac{f_{class,arg} + \beta}{f_{class,\cdot} + N\beta} \frac{f_{class,vb} + \alpha_{class}}{f_{\cdot,vb} + \sum_{class'} \alpha_{class'}} \tag{3.4}$$

An important observation is that **Classical LDA** helps calculate the probability of an argument conditioned on the verb and grammatical relation, rather than the joint probability. As I am testing the performance of the model on a joint probability gold standard, I needed to multiply the probability given by the model by the relative frequency of the verb within that grammatical relation in the training corpus, according to equation 3.1.

Rooth LDA

Rooth LDA, however, follows a slightly different the generative story, which is based on the intuition that verb-argument pairs are generated together from the same class-specific distribution. Unlike in Classical LDA, where the argument is generated from a verb-dependent class, the verb-argument interaction is modelled in a symmetric fashion in Rooth LDA, as the three-step generative story taken from from (Ó Séaghdha, 2010) illustrates:

1. For each relation rel, a multinomial distribution over interaction classes is drawn from a Dirichlet distribution with parameters α.

2. For each interaction class, a multinomial distribution over arguments is drawn from a Dirichlet distribution with parameters β, and a multinomial distribution over verbs is drawn from a Dirichlet distribution with parameters γ.

3. A verb-argument pair is generated by first drawing an interaction class, and then by independently drawing an argument and a verb from the argument distribution and the verb distribution of the class, respectively.

Equation 3.5 estimates the plausibility of a verb-argument pair in the context of a relation, once we have trained the parameters of the model. The equation uses the same notations as equation 3.3 for Classical LDA, with the newly introduced variable V to denote the size of the verb vocabulary.

$$P(arg, vb|rel) = \sum_{class} P(arg|class)P(vb|class)P(class|rel) \tag{3.5}$$

$$\propto \sum_{class} \frac{f_{class,arg} + \beta}{f_{class,\cdot} + N\beta} \frac{f_{class,vb} + \gamma}{f_{class,\cdot} + V\gamma} \frac{f_{class,rel} + \alpha_{class}}{f_{\cdot,rel} + \sum_{class'} \alpha_{class'}} \tag{3.6}$$

Another notable different between Rooth LDA and Classical LDA is that Rooth LDA helps calculate the joint verb-argument probability, which I could directly correlate to the gold standard probabilities.

3.2 Cross-lingual transfer models

Unlike the algorithms presented in section 3.1, which provide solutions for predicting selectional preferences in languages where dependency-parsed resources are available, the algorithms described in sections 3.2.1 and 3.2.2 are concerned with effectively transferring information about selectional preferences from resource-rich languages into resource-poor languages. While previous research by Peirsman and Pado indicates that it is possible to automatically infer plausibilities across languages using bilingual vector spaces, I use the algorithms described in this section to both verify their results, and to determine the extent to which cross-lingual transfer can be more effective than the monolingual methods.

3.2.1 Translation based on dictionary resources

In order to gauge the potential of cross-language transfer of selectional preferences, the first method I employed was to compile a list of translations into English of the verbs and arguments from the German, Spanish and Romanian test datasets, and to approximate the plausibility of non-English tuples by the state of the art monolingual plausibility of the tuple translations into English, as previously indicated in equation 2.3 from section 2.3. This experiment was designed based on the intuition that running the tests with high-quality translations would yield an upper bound on cross-lingual performance.

i experimented with two methods of obtaining English translations of the verbs and arguments in other languages, both of them easily available at no cost for typically resource-poor languages. The first method was to use publicly-available online dictionaries, and where more translation alternatives are provided, to pick the one with the highest language frequency. As I am interested in the applicability of my methods to resource-poor languages, I did not rely on an actual corpus to assess the language frequency of the translations, but rather used the number of page hits in a search query to Google as an approximation. The second method of obtaining automatic translations I tried was inputting the individual words into Google Translate[1].

While both methods yielded correct translations for the verbs and arguments in the test datasets, I observed a severe shortfall. In many of the cases where the translation from a foreign language into English is not a one-to-one mapping, an out-of-context translation was picked, because the verb and the noun were translated independently of each other. For example, the German tuple {*kürzen*, *Leistung*, *direct object*} was translated to {*shorten*, *performance*, *direct object*}, which is inaccurate. In order to address this issue, the third and final method of hardcoding the translations I used was to obtain them from a native speaker of each of the three foreign languages. In cases where the native speaker provided alternative translations for the

[1] http://translate.google.com/

same verb across different tuples, I only retained the most frequent one. This is because I wanted to keep the experiment as an idealized version of the algorithm described in the following section.

3.2.2 Translation with bilingual vector spaces

The algorithm I studied for the automatic cross-language transfer of selectional preferences was first introduced in (Peirsman and Padó, 2010). Unlike the benchmarking method presented in section 3.2.1, which relies on high-quality translation dictionaries, this algorithm only requires large, unparsed corpora in the source and target languages, provided that a method for computing plausibilities in the resource-rich target language is already available. Furthermore, according to the claims in the original paper, it is not necessary for the corpora to be related for the algorithm to work.

The main idea behind the algorithm is to build a common vector space in which to represent the nouns and verbs of both the source and target languages. The coordinates of a given word in the vector space are computed based on that word's co-occurrence statistics with the word labels on the axes, where co-occurrence statistics are derived from the unparsed corpus for the language of the represented word. Furthermore, in order to be able to build a unified, bilingual vector space, the word labels on the axes must be translation pairs of each other in the two languages. Figure 3.1 illustrates a conceptual example of a bilingual vector space for English and German.

The intuition behind using bilingual vector spaces is that words which are mutual translations of each other will show similar co-occurrence patterns, and can be identified based on the distance between their corresponding vectors. The quality of the translations obtained with this method depends on two aspects: the axes of the vector space, and the co-occurrence statistics used. On one hand, a larger number of axes labelled with correct translation pairs will give more discriminative power to the vector space, while also having two disadvantages: the time complexity of the search for translation candidates is linear in the number of dimensions, and translation pairs con-

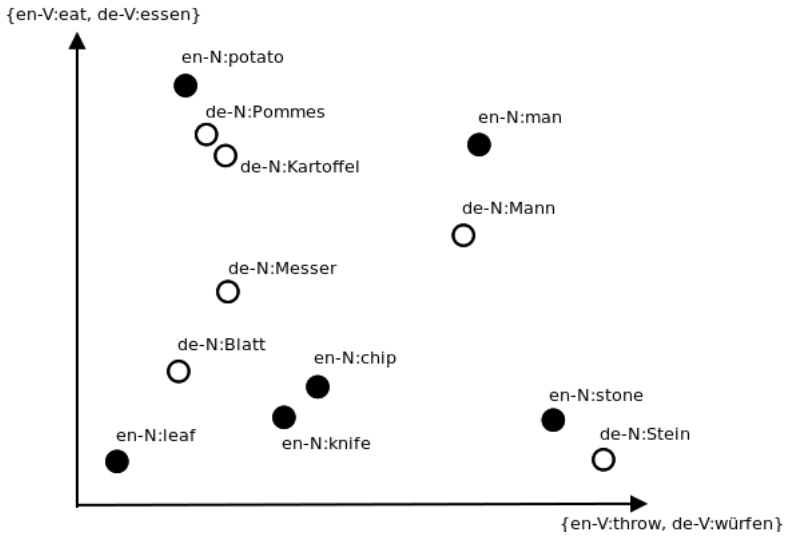


Figure 3.1: Concept illustration of a bilingual vector space for English (black dots) and German (white dots), with two verb axes, and 12 noun words.

stitute a costly input resource. On the other hand, since parsed corpora are not used, only POS-tag and word order information can be used for computing the co-occurrence statistics, with the original authors suggesting context-window frequency counts.

3.3 Summary

In this chapter, I described the selectional preference estimation algorithms implemented and tested within this project, and showed what steps and data resources are necessary for training each of them. The algorithms discussed fall into two classes: monolingual models, and cross-lingual transfer models. While the former require dependency-parsed data for training, the latter can be trained with dictionary resources and/or POS-tagged data to transfer to a resource-poor source language in which dependency-parsed data would not normally be available, the plausibilities estimated using a monolingual model in a target language.

Chapter 4

Data set and resources

4.1 Overview of the corpora used

For the purpose of training the plausibility prediction algorithms described in Chapter 3, I needed corpora in each of the four languages studied: English, German, Spanish and Romanian. The motivations behind my choice of these four languages are diverse. English, while being the language with the largest amount of available data resources, is also the first and only language in which latent variable models were applied to the estimation of selectional preferences, as previously described in (Ó Séaghdha, 2010). While German and Spanish cannot be classified as resource-poor languages, choosing them ensured me access to accurate parsers and sufficient data, allowed me to study both a Romanace, lexically-isolating language, and a Germanic, lexically-agglutinative language, and enabled a direct comparison to the results of previous research in cross-lingual selectional preference learning described in (Peirsman and Padó, 2010). Finally, choosing Romanian allowed me to test my methods on a genuinely resource-poor language, which has no available large human-annotated corpus or a widely-known dependency parser.

There are two types of information I extracted from the four corpora: syntactic information, for which it was necessary to have access to a dependency-

parsed version of the corpus data, and word-window context information, for which a tagged version of the corpus data was sufficient. Additionally, I also extracted translation pairs from both dictionary copora, and from corpora that have been aligned at the document-level.

For English, I based my studies on the British National Corpus. For German, I used the TIGER Corpus, and for Spanish I used the AnCora Corpus. I could find no dependency-parsed available corpora in Romanian. However, for each of the four languages I also used the respective Wikipedia and Wiktionary dumps as additional, unparsed corpora. Sections 4.2 to 4.4 describe the corpora and the necessary data preparation steps in more detail.

4.2 Using Wikipedia as a corpus

Gaining access to large corpora on a *variety of topics* is a difficult task even for languages which are not traditionally considered resource-poor in NLP, such as Spanish or German. Thus, the main advantage of Wikipedia is that it can be downloaded and used freely under the Creative Commons License. With seven languages (including English, German and Spanish) at over $1,000,000$ articles, and an additional 38 languages (including Romanian) with between $100,000$ and $1,000,000$ articles as of May 2013, Wikipedia is the most easily accessible source of well-formed, general-purpose text for many European and Asian languages.

The Wikipedia database dumps available for download are compiled monthly and published online[1] for each of the supported languages. The articles dump, containing only the latest revision of the source code of all articles, is published in a single, compressed XML file. I parsed this file to retrieve the contents of the individual articles, stored in the <page> nodes of the document. The title of the articles can be retrieved from the <title> child of the corresponding <page> node, while the Textile source code for the article is stored in the <text> child. In order to obtain the plain-text version of the

[1] http://en.wikipedia.org/wiki/Wikipedia:Database_download

article contents, I stripped away the markup language from the extracted source code using the WikiCloth[2] and Nokogiri[3] Ruby libraries. Further pre-processing steps on the output text articles included removing tables, images, scripts, the inter-language links, the edit links, thumbnails, and the references.

There are two main types of Wikipedia articles: *topic articles*, which describe an entity (eg. the article for *"Music"*[4]), and *list articles*, which contain an enummertation of entities (eg. the article for *"List of Mountains in Europe"*[5]). For the purpose of compiling my corpora, I discarded all *list articles*, as well as all articles which were under 10 KB in size, and grouped the remaining *topic articles* in shards of $5,000$ articles each, to make them easier to process.

4.3 Corpora description and preparation by language

4.3.1 English

The **British National Corpus** (BNC) is a 100-million word corpus containing text samples that are typical of current British English. The BNC is a general-purpose corpus, containing written samples, as well as transcripts of spoken language samples, on a variety of topics.

For the purpose of my monolingual experiments, I used an XML-encoded, dependency-parsed version of the BNC corpus that was obtained with RASP. The RASP Parser, described in more detail in (Briscoe et al., 2006), encodes Grammatical Relations (GRs) directly in the XML output. However, two additional processing steps were required for the accurate extraction of the

[2]http://code.google.com/p/wikicloth/
[3]http://nokogiri.org/
[4]https://en.wikipedia.org/wiki/Music
[5]http://en.wikipedia.org/wiki/Mountains_in_Europe#Europe

verb-argument pairs: I had to re-classify the subjects of verbs in the passive voice as direct objects, and I had to distribute grammatical relations over coordinated dependents. Furthermore, I discarded all verb-argument pairs where the argument is not tagged as a common noun, as well as all lemmas containing characters that are neither unicode letters, nor the hyphen character.

For the purpose of building vector spaces for my cross-lingual experiments, however, I started out from the **English Wikipedia**, because it is related to the Wikipedias in other languages in terms of content. The article database snapshot that I used was the one from March 2013, and after applying the pre-processing steps described in section 4.2, I was left with 299, 632 articles, containing a total of 569, 445, 316 tokens. I tagged this plain-text output using the TreeTagger, described in more detail in (Schmid, 1994), and later in (Schmid, 1995). The tagging model that I used was trained on the PENN Treebank, which uses the associated PENN Treebank tagset, described in (Santorini, 1990).

4.3.2 German

The first corpus I used for my German monolingual experiments was the **TIGER Corpus**. The corpus and its annotation schema were described in detail in (Brants et al., 2002) and in (Brants et al., 2004). There are two main versions of the TIGER corpus, with the second version having two releases. As TIGER 2.2 is only a cleaned-up version of the 2.1 release, I used version TIGER 2.1 as the starting point for my monolingual experiments.

In order to address the data sparsity caused by the relatively small size of TIGER, I turned to the **German Wikipedia** as an alternative corpus, which I used in both POS-tagged and dependency-parsed format. I used the March 2013 database dump of the German Wikipedia, from which I extracted 105, 820 articles, containing rougly 198, 763, 417 tokens, by applying the pre-processing described in section 4.2.

For building the vector spaces necessary in the cross-lingual experiments, I tagged the plain-text output using the TreeTagger. The tagging model that I used with TreeTagger was trained on the Negra Corpus, and it used the associated Stuttgard-Tübingen Tagset (STTS), described in detail in (Schiller et al., 1995). However, in order to use Wikipedia as an altenative corpus to TIGER, it was also necessary to parse it and extract syntactic information. Given the relatively free word order in German and the fact that constituency trees cannot be trivially translated to dependency relations, there are currently very few parsers available for this language which can output dependencies in the Stanford format with accuracy comparable to that for English or Spanish. I used the Zurich Dependency Parser for German (also known as ParZu, but formerly known as Pro3GresDE), which is able to output dependencies in the widely-known CoNLL format. The parser is described in (Sennrich et al., 2009), where it is reported to outperform both MaltParser and MSTParser in the prediction of subjects and objects.

4.3.3 Spanish

For Spanish, the first corpus that I used was version 2.0 of the **AnCora-ESP Corpus**, which is distributed in CoNLL format. AnCora stands for *'Annotated Corpora'*, and it is one of the most frequently-cited Spanish corpora in NLP publications. More details about the corpus and its annotation guidelines are given in (Martí et al., 2007) and (Taulé et al., 2008).

Again, I turned to the **Spanish Wikipedia** as an alternative, larger-sized corpus. As in the case of the German Wikipedia, I used the March 2013 database dump, from which I was able to extract 72,764 articles following the aplication of the pre-processing steps. The extracted articles amassed a total of 153,966,362 tokens, about one and a half times the size of the BNC.

As in the case of the German Wikipedia, I used both the POS-tagged and the dependency-parsed versions of the corpus. For the purpose of building the vector spaces used in the cross-lingual experiments, I tagged the output using the TreeTagger with a tagging model trained on the Spanish CRATER

Corpus, and using a simplified version of the original CRATER tagset, while for the purpose of extracting syntactic information, I chose to parse it with Malt Parser, which was described in (Nivre et al., 2006) and (Nivre et al., 2007), and which outputs Stanford dependencies in the CoNLL format. I used version 1.7.2 of Malt running the Nivre algorithm, with a parse model trained on the IULA Treebank, described in (Marimon et al., 2012). One notable difference between Malt and the parsers I used for German and Romanian is that the input to Malt is required to be in CoNLL format, and tagged using FreeLing tag set (adapted from the Eagles tag set). I used the tagger component of FreeLing 3.0, described in (Padró and Stanilovsky, 2012), and then converted its output to the CoNLL format.

4.3.4 Romanian

Unlike for English, Spanish, or German, no freely available, dependency-parsed corpora exist for the Romanian language. Thus, I had to rely on the **Romanian Wikipedia** as my only corpus of well-formed text. I expect this to be the common scenario for resource-poor languages, for which well-formed samples of the language should still be easy to obtain, even though annotated data might be unavailable. Following the application of the pre-processing steps described in section 4.2 on the database snapshot of the Romanian Wikipedia from March 2013, I was left with $9,668$ articles amassing a total of $17,834,631$ tokens, which is still comparable to the size of the human-annotated German and Spanish corpora.

Similarly to the other Wikipedias, I used the Romanian corpus in two forms: POS-tagged only, in order to build a vector space for my cross-language experiments, and fully parsed in order to extract syntactic information for the monolingual algorithms.

In order to compile the POS-tagged version of the corpus, I used a recently-developed hybrid POS-tagger which combines statistic tagging models and a rule-based system. The tag set used by this tagger is a simplified version of the RACAI tag set used in the training data, called the MSD tag set. A

more detailed description of the MSD tag set is given in (Toutanova et al., 2003). The tagger achieves 96.6% precision on the *'1984'* corpus. Further details about its design and performance are given in (Simionescu, 2011).

In order to parse the data, I used the UAIC Romanian FDG Parser, which is as of the moment of writing this paper, the only publicly available dependency parser for the language. The parser is accessible onlina via a web service[6], and it produces output in XML format, which I post-processed and converted to CoNLL format. Note that under normal circumstances, I would not expect a dependency parser to be available for a resource-poor language, but I used syntactic information for Romanian with the experimental purpose of determining how well the cross-lingual methods compare to the monolingual ones under conditions of data scarcity.

4.4 Translation dictionaries

In order to make cross-language plausibility estimation possible, in addition to the vector space for each language built from the POS-tagged corpora, I also need a corpus of translations for aligning them. I expect that even for resource-poor languages, translation dictionaries should be a readily-available and inexpensive resource. In sections 4.4.1 and 4.4.2, I describe two of the methods of getting translations that I explored.

4.4.1 Wikipedia inter-Language links

Using the human-annotated inter-language links present in Wikipedia topic articles is a straightforward way of obtaining translation pairs based on which to align the vector spaces of different languages. The inter-language links between the articles are easily parsed from the XML article databse dumps: the links are indexed by the two-letter language code of the foreign languages which contain articles on the same topic, and provide alternative translations

[6] http://nlptools.info.uaic.ro/WebFdgRo/

to the article title. *'Foreign languages'*, in this context, are taken to mean languages different from the language of the database dump being parsed. I parsed the English database dump for extracting translation pairs, because it is the largest in number of articles and the most comprehensive of the Wikipedias, and I discarded the translation pairs containing multiple words or characters besides letters and hyphens.

Although easy to extract, using the inter-language links as sets of translation pairs has a number of shortfalls. One of the problems is that not all article titles are made of only one word. Even among one-word titles, a large number of them are proper nouns, or otherwise infrequent terms, and are unsuitable for aligning vector spaces. Another problem is that virtually all of the one-word article titles are nouns, while bilingual vector spaces require a balanced verb-noun translation set to give accurate similarity estimates. A third and final problem is that the inter-language links are article titles, and are thus formatted in title casing across all languages. While not a problem for German, the title casing causes problems when trying to distinguish between common nouns and proper nouns across the translation tuples, and when trying to normalize the casing of the links.

4.4.2 Wiktionary

Another alternative for easily obtaining translation pairs between the four languages using only publicly-available data is to use Wiktionary[7]. Wiktionary is a collaborative project that was designed and licensed on the same principles as Wikipedia, and it serves as a free multilingual dictionary. Unlike the inter-language links taken from Wikipedia, which are neither necessarily one-word expressions, nor provide translations for other parts of speech besides nouns, the contents of Wiktionary comprise a true dictionary. Furthermore, Wiktionary contains alternative translations for both verbs and nouns in places where the alignment between languages is not one-to-one. Where more than one translation was provided, I only considered the top

[7] http://en.wiktionary.org/wiki/Wiktionary:Main_Page

two most frequent translations.

The Wiktionary database dumps are freely available online for download[8]. Following the same argument as in the case of the Wikipedia inter-language link extraction, I used the English Wiktionary dump from April 2013 for maximized coverage of the extracted translations.

4.5 Summary of resources used

In this chapter, I described the corpora I used for the four languages, justified the use of Wikipedia as an easily-accessible corpus of well-formed text in a given language, and described how to extract translation pairs from Wikipedia and Wiktionary, showing that freely-available text resources can be used to build parsed corpora. Tables 4.1 and 4.2 below summarize the amount of information I extracted from each corpus.

Training Corpus		Size (tokens)	Parser	Verb-Direct Object		Verb-Subject	
				Verbs	Pairs	Verbs	Pairs
English	BNC	100,000,000	RASP	8,309	2,702,367	9,443	3,912,915
German	TIGER	888,581	ParZu	2,602	23,348	2,290	36,939
	Wikipedia	198,763,417	ParZu	34,262	3,855,209	41,920	7,242,897
Spanish	AnCora	500,000	manual*	1,385	14,795	1,462	15,140
	Wikipedia	153,966,362	MaltParser	38,019	4,802,171	21,567	2,858,669
Romanian	Wikipedia	17,834,631	UAIC FDG	11,192	540,909	8,958	470,444

* Parser-aided.

Table 4.1: A summary of all the dependency-parsed corpora used within the project. Only common nouns were included in the verb argument counts.

Source	English - German		English - Spanish		English - Romanian	
	Verbs	Nouns	Verbs	Nouns	Verbs	Nouns
Wikipedia inter-language links	0	134,866	0	113,509	0	34,542
Wiktionary	2,968	15,002	3,026	12,912	1,163	5,305

Table 4.2: A summary of all the translation pairs used within the project.

[8]http://dumps.wikimedia.org/backup-index.html

Chapter 5

Experiments

5.1 Testing methodology

The qualitative evaluation of all monolingual and cross-lingual models of selectional preference was performed by correlating the predicted plausibilities with human judgements of joint verb-argument probabilities on datasets of verb-subject and verb-direct object tuples. Although I used joint probabilities, rather than conditional probabilities, the two are interdependent under the constraints of a language model, as demonstrated in equation 5.1.

$$P(vb, arg, rel) = P(arg|vb, rel) \cdot P(vb, rel) \tag{5.1}$$

This evaluation scheme was previously used in both (Peirsman and Padó, 2010) and (Ó Séaghdha, 2010), enabling a direct comparison of my results to those reported in previous research.

I evaluated correlation by computing the Spearman ρ correlation coefficient, also known as *Spearman's rank correlation coefficient*, which measures how well the dependence relationship between two variables, X and Y, can be described using a monotonic function taking values between -1 and $+1$ (with 0 indicating lack of correlation). In my case, the variables X and Y are

the vectors of gold standard and predicted plausibilities, respectively. The formula for the coefficient is given by equation 5.2 below, where the values for X and Y are converted to ranks x and y (the ranks for equal values are taken to be the mean of their actual ranks).

$$\rho = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5.2}$$

It should be noted that since defaulting to zero-plausibility on test tuples which contain unseen verbs and/or arguments has the result of distorting the Spearman's ρ, on certain experiments I only report correlation on a sub-set of the test dataset. In such cases, I also report the coverage of the evaluated method.

5.2 Test datasets by language

Sections 5.2.1 to 5.2.4 below describe the gold standard dataset for each of the four languages in further detail.

5.2.1 English

In the case of English, I used two datasets for testing.

The fist dataset I used was taken from previous work on the modelling of human plausibility judgements, described in (Padó et al., 2007). The data set contains human judgements on 173 verb-direct object pairs, obtained through the combination of 15 verbs with 11 or 12 arguments, and on 202 verb-subject pairs, obtained through the combination of 18 verbs with 11 or 12 arguments. The plausibility judgements in the dataset are on a scale from 1.0 to 7.0, with low scores corresponding to unlikely combinations, and high scores corresponding to likely ones. According to the original authors, the verb-argument combinations are sampled from corpus co-occurrences derived

from the Penn Treebank and the FrameNet Corpus using a uniform, rather than bimodal distribution. This dataset modelling approach is different in principle from the previous psycholinguistic experiments on language constraints, published in (McRae et al., 1998), but it leads to a more realistic task, as acknowledged in both (Padó et al., 2007), and (Ó Séaghdha, 2010).

The second dataset I used was initially compiled by Keller and Lapata, and published in (Keller and Lapata, 2003). It contains human judgements on 180 verb-direct object pairs, which resulted from the combination of 60 verbs with three arguments each. Unlike the first English dataset I used, where the arguments had been uniformly sampled from the corpora, the three arguments in this dataset were sampled from high-, medium- and low-frequency combinations from both the British National Corpus and the NANTC Corpus. Further differences between the two datasets include the fact that the Keller and Lapata dataset lacks human judgments for verb-subject combinations, and uses a different method for computing plausibility scores, which are on a logarithmic scale.

By using these two datasets, I could verify the results cited in (Ó Séaghdha, 2010) on the use of generative models for predicting selectional preferences.

5.2.2 German

For German, I used the same test dataset that was cited in (Peirsman and Padó, 2010), in order to enable a direct comparison with their results on the cross-language induction of selectional preferences. The verb-argument tuples were initially collected and presented in (Brockmann and Lapata, 2003), which also gives the gold standard of human-elicited judgements and a detailed description of the methodology used for assembling similar gold standard datasets for other languages.

The dataset contains a total of 90 verb-argument-relation triples, with 30 triples covering each of the direct object, subject and oblique object grammatical relations. The triples in the dataset were sampled from the German

Süddeutsche Zeitung, which contains a total of 179 million words of newspaper text, after discarding corpus words with less than one one-per-million occurrences. For the purpose of my experiments, I only used the direct object and subject sets, which are composed of ten verbs paired with three arguments each. The arguments were divided into three corpus frequency bands: high, medium and low, and an argument was randomly chosen from each band. This sampling strategy is similar to the one used by Keller and Lapata in building their English language dataset, as it ensures the equal representation of likely and unlikely verb-argument combinations; equal representation is considered important by the authors in (Brockmann and Lapata, 2003) for accurately evaluating various models under conditions of both high and low corpus counts.

5.2.3 Spanish

For Spanish, I chose to use the same test dataset that was cited in (Peirsman and Padó, 2010), following the same reasoning I applied for the German dataset: my choice allowed me to make a direct comparison with their results on the cross-language experiments. Unlike German, however, where the plausibility dataset had been taken over from the previous study in (Brockmann and Lapata, 2003), the triples along with their human judgements for the Spanish gold standard were compiled by Peirsman and Pado themselves, in the absence of previously published gold standards for the language.

The data set contains 60 verb-argument triples, with 30 triples covering the direct object and 30 triples covering the subject grammatical relations. The triples were sampled from two Spanish corpora: the relatively small, $500,000$ words AnCora corpus, which contains news text, and the much larger, 18 million words Encarta corpus, which contains encyclopaedic text. As indicated by the original authors, the sampling was done according to the principles described in (Brockmann and Lapata, 2003). The arguments were extracted from the corpora and divided into three frequency bands after discarding rare arguments, and one argument was randomly chosen from each

band. Unlike the German dataset, the human judgements were elicited on a five-point Likert scale using the Amazon Mechanical Turk crowdsourcing platform. The scoring method for the gold standard was proven by the authors to be equivalent to the method described in (Brockmann and Lapata, 2003) by re-assessing the German dataset using the new methodology, and proving a Spearman ρ correlation of approximately 0.9.

5.2.4 Romanian

The main reason why I chose Romanian as one of the languages for our study is because Romanian is a truly resource-poor language. As acknowledged in (Peirsman and Padó, 2010) by the original authors, the use of Spanish and German as resource-poor languages simulated by varying the amount of corpus data used. Furthermore, unlike English, German, or Spanish, none of the algorithms described in chapter 3 had previously been tested on Romanian, and thus there was no test dataset and no gold standard of human plausibility judgments previously available for evaluating the applicability of these models to the language.

Thus, starting from the principles described by (Brockmann and Lapata, 2003), I created a new test dataset for Romanian. I used the Romanian Wikipedia, described in section 4.3.4, as a corpus from which to sample verbs and arguments. I uniformly sampled ten verbs, and then for each of the sampled verbs, I divided the corpus arguments into three frequency bands and uniformly sampled one argument from each band. I used the same verbs for both grammatical relations considered, *direct object* and *subject*. At 30 triples per grammatical relation, the size of the resulting dataset is the same as the size of the test datasets for German and Spanish.

In order to elicit human plausibility judgements on the dataset, I created two online questionnaires, one for each grammatical relation. Each questionnaire was made up of 30 multiple-choice questions, one question per verb-argument pair, asking the respondent to judge the joint plausibility of a verb-argument pair on a five-point Likert scale, with values ranging from 1.00, which means

Table 5.1: The gold standard plausbility judgements for the Romanian dataset. English translations are given below each word, in italic.

Verb	Direct Objects		
	High Frequency	Medium Frequency	Low Frequency
a scrie	**articol** (4.82)	**rând** (4.00)	**indicație** (3.50)
to write	*article*	*line*	*guideline*
a traversa	**stradă** (4.97)	**perioadă** (4.97)	**podiș** (3.28)
to cross/to pull through	*street*	*time period*	*plateau*
a lovi	**om** (4.20)	**braț** (2.94)	**cadru** (2.28)
to hit	*human/person*	*arm*	*frame*
a închide	**ochi** (4.85)	**poartă** (4.82)	**frizerie** (3.34)
to close	*eye*	*gate*	*barber's shop*
a lega	**șnur** (4.37)	**fular** (3.40)	**snop** (3.17)
to tie	*string*	*scarf*	*bundle*
a finanța	**proiect** (4.85)	**studiu** (4.14)	**timp** (1.82)
to finance	*project*	*study*	*time*
a cuceri	**lume** (4.22)	**problemă** (1.98)	**virus** (1.57)
to conquer	*world*	*problem*	*virus*
a amplifica	**problemă** (3.45)	**efect** (2.94)	**alarmă** (2.48)
to amplify	*problem*	*effect*	*alarm*
a combate	**discriminare** (4.48)	**tendință** (3.42)	**impuls** (2.68)
to battle	*discrimination*	*tendency*	*impulse*
a colora	**peisaj** (3.57)	**papion** (1.51)	**rebut** (1.37)
to colour	*landscape*	*bow tie*	*scrap*

Verb	Subjects		
	High Frequency	Medium Frequency	Low Frequency
a scrie	**redactor** (4.65)	**stilou** (4.61)	**cercetător** (3.86)
to write	*editor*	*pen*	*researcher*
a traversa	**călător** (3.63)	**râu** (3.81)	**pacient** (2.25)
to cross/to pull through	*traveller*	*river*	*patient*
a lovi	**om** (3.52)	**pumn** (4.31)	**întâmplare** (1.72)
to hit	*human/person*	*fist*	*incident*
a închide	**paznic** (3.61)	**magazin** (4.36)	**cutie** (3.75)
to close	*watchman*	*store*	*box*
a lega	**muncitor** (2.75)	**tunel** (3.00)	**jurământ** (3.50)
to tie	*worker*	*tunnel*	*oath*
a finanța	**guvern** (4.52)	**proiect** (4.27)	**inițiativă** (3.50)
to finance	*government*	*project*	*incentive*
a cuceri	**inamic** (3.97)	**sportiv** (3.84)	**melodie** (3.04)
to conquer	*enemy*	*sportsman*	*song*
a amplifica	**stație** (4.18)	**condiție** (2.68)	**comentariu** (2.25)
to amplify	*station*	*condition*	*comment*
a combate	**poliție** (4.00)	**doctor** (3.36)	**medicament** (4.34)
to battle	*police*	*doctor*	*medicine*
a colora	**vopsea** (4.38)	**fiu** (2.40)	**cafea** (2.13)
to colour	*paint*	*son*	*coffee*

"unusual" to 5.00, which means *"typical"*. The order of the questions was randomized among respondents, but all respondents were asked to rate the entire set of the verb-argument pairs.

I then used a popular social network to collect human judgements on the two sets of triples from college-educated native speakers of Romanian. The respondents were told that the purpose of the survey is to test Natural Language Processing techniques on the Romanian language, and no rewards were given in return for answering the questionnaires. The choice of whether to answer either both questionnaires, or just one of them, was left entirely to the respondents. I left the survey open for 48 hours, and I was able to gather 35 judgements for each verb-direct object pair, and 44 judgements for each verb-subject pair. I then aggregated the judgements by taking the mean for each triple, a technique proven in (Peirsman and Padó, 2010) to be equally reliable to the more complicated Magnitude Estimation method originally described in (Brockmann and Lapata, 2003). The scale of the resulting gold standard plausibility judgements is linear.

The dataset, along with the gold standard plausibility judgements, is given for reference in table 5.1.

5.3 Monolingual results

5.3.1 Baselines

The performance of the corpus frequency monolingual baseline algorithm for *direct object* and *subject* grammatical relations, on all parsed training corpora available, is listed in table 5.2 above. Note that for English, two results are given for the British National Corpus, one for each test dataset described in section 5.2.1. Also note that for the smaller-sized corpora, TIGER and AnCora, correlation results are only reported for the subset of the test dataset which has coverage in these corpora. For AnCora, this way of reporting correlation is taken over from (Peirsman and Padó, 2010), who used it to

Training Corpus		Direct Object		Subject	
		ρ	Coverage	ρ	Coverage
English	BNC (Pado)	**0.614**	100%	**0.196**	100%
	BNC (K&L)	**0.683**	100%	-	-
German	TIGER	-0.029	80%	0.286	80%
	Wikipedia	**0.582**	100%	**0.417**	100%
Spanish	AnCora	-0.304	60%	0.000	80%
	Wikipedia	**0.442**	100%	**0.325**	100%
Romanian	Wikipedia	**0.639**	100%	**0.193**	100%

Table 5.2: The performance of the corpus frequency monolingual baseline.

report their results for the similarity-based monolingual baseline on the same corpus. Although no results are published in (Peirsman and Padó, 2010) for the TIGER corpus, I chose the same methodology of reporting correlation across all algorithms tested, based on the observation made in section 5.1 that the r, and especially the ρ coefficients are susceptible to distortion from zero-values.

The results of this experiment indicate that, in the presence of roughly the same amount of parsed input data, selectional preferences are harder to model for *subject* relations than they are for *direct object* relations; all of the top-performing results highlighted in bold font for each language were derived from training corpora of between 3 and 8 million verb-argument-relation input triples, with the notable exception of Romanian, which had on the order of half a million triples for each relation.

The results also help confirm my hypothesis that the test datasets for different languages are at **different levels of difficulty**. While this aspect was also mentioned in (Peirsman and Padó, 2010), they only investigated the difficulty of the German dataset, and did so through comparison with ontology-based models of selectional preference, which do not allow an objective comparison between languages due to the nature of the ontological resources employed. With my approach, the correlation coefficients highlighted in table 5.2 show that in the case of *direct objects*, the Spanish dataset is consistently harder to model than the German one, while the English and Romanian datasets are

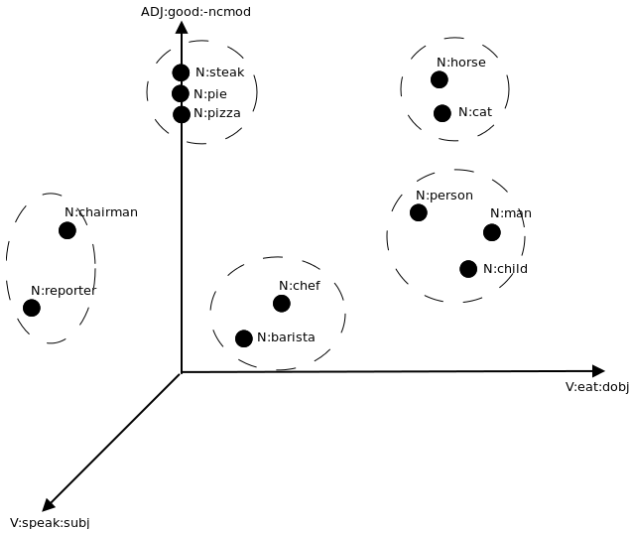


Figure 5.1: Concept illustration of a syntactic context vector space for English, with three grammatical relation-lemma axes, and 12 arguments. The dotted ellipses highlight closed groups of nearest neighbours, showing semantic similarity.

the easiest ones to model. However, in the case of *subjects*, while Spanish still proved harder to model than German, the English and Romanian datasets proved very difficult to model, falling to the bottom of the ranking with correlations below 0.200.

Following the advice for compiling monolingual baselines given in (Peirsman and Padó, 2010), I implemented and optimized a similarity-based algorithm, built on the description we gave in section 3.1.2. In the implementation of the method's formula, I used the verb-argument co-occurrence matrices used for the corpus frequency baseline for computing relative frequencies. For computing a similarity metric, I built a syntactic context vector space from the dependency-parsed corpora in which I represented all argument heads, and used the classical cosine distance to measure similarity between arguments.

A syntactic context vector space is a vector space whose axes are labelled with {*grammatical relation, lemma*} pairs. Figure 5.1 conceptually illustrates a syntactic context vector space. Both the grammatical relations in which

the argument is the head, and those in which it is the dependent were taken into consideration when building the vector space, and all relation types were allowed. Relations in which the argument is the dependent are prefixed with a '−' sign in the figure.

For optimizing the vector space, I indirectly measured its performance against the final selectional preference correlation coefficients, and concluded that the best results are obtained when allowing on the axes only lemmas belonging to the following parts of speech: *noun*, *verb*, *adjective*, *adverb*, *preposition*. While my experimental observations conformed to my expectations that open class parts of speech should be informative in the vector space, and that closed class parts of speech shouldn't, I discovered that the inclusion of prepositions leads to more accurate syntactic similarity measures. This is because relations to prepositions reflect selectional preferences for oblique objects and noun modifiers, both of which are informative for computing syntactic similarity metrics.

I followed the original specifications from (Peirsman and Padó, 2010) in choosing pointwise mutual information as the co-occurrence statistic for my vector space, although I observed only a small performance difference between PMI and normalized frequency counts. I also followed the original specifications in choosing only the top $2,000$ most frequent $\{$*grammatical relation, lemma*$\}$ pairs as dimensions.

Training Corpus		Direct Object		Subject	
		ρ	Coverage	ρ	Coverage
English	BNC (Pado)	0.519	100%	0.050	100%
	BNC (K&L)	0.598	100%	-	-
German	TIGER	**0.058**	80%	0.124	80%
	Wikipedia	0.496	100%	0.415	100%
Spanish	AnCora	**0.188**	60%	**0.179**	80%
	Wikipedia	0.485	100%	0.362	100%
Romanian	Wikipedia	0.566	100%	-0.008	100%

Table 5.3: The performance of the similarity-based monolingual baseline. Highlighted values indicate better performance than that of the frequency count baseline.

The results obtained with resulting optimized vector spaces are summarized in table 5.3. The correlation coefficients reported were lower than those of the corpus frequency baseline for the large corpora. However, for the smaller AnCora and TIGER corpora, the performance was better than that of the baseline in all cases except German *subjects*. My explanation for this phenomenon is that similarity-based methods rely on over-generalization, which is quickly overtaken in terms of performance by frequency estimates when increasing the amount of training data. Despite the differences between the two methods, the results still verify my observation that selectional preferences are harder to predict for subjects than for direct objects.

5.3.2 Latent Dirichlet Allocation

I used the verb-argument co-occurrence matrices extracted from the corpora for the frequency count baseline algorithm, and used the MALLET toolkit to train the latent variable models on them. Following the methodology specified in (Ó Séaghdha, 2010), I used topic models of 100 classes, which I trained for 2000 iterations. Using the same methodology enabled me to make a direct comparison to the results reported on the Keller and Lapata English test dataset. I uniformly initialized the α, β and γ hyperparameters to 0.01, and then re-sampled them every 50 iterations following the 200-iterations burn-in period, according to the detault settings of the MALLET toolkit. However, although the results I report were based on 2000 topic model training iterations, I was unable to measure any performance differences between training for 500, 1000, or 2000 iterations. Furthermore, unlike the approach taken in (Ó Séaghdha, 2010), I trained the Rooth LDA model using the same number of classes and for the same number of iterations as in the case of the Classical LDA model.

Table 5.4 summarizes the performance measurements of my LDA implementations. The only performance comparison I can make between my results and those reported in (Ó Séaghdha, 2010) is for the Keller and Lapata *direct object* test dataset, as the original paper did not study verb-subject inter-

Training Corpus		Direct Object		Subject	
		ρ	Coverage	ρ	Coverage
Classical LDA					
English	BNC (Pado)	0.602	100%	**0.233**	100%
	BNC (K&L)	**0.708**	100%	-	-
German	TIGER	0.008	80%	**0.305**	80%
	Wikipedia	0.451	100%	**0.469**	100%
Spanish	AnCora	0.287	60%	**0.386**	80%
	Wikipedia	0.437	100%	**0.463**	100%
Romanian	Wikipedia	0.607	100%	**0.258**	100%
Rooth LDA					
English	BNC (Pado)	0.603	100%	**0.235**	100%
	BNC (K&L)	**0.711**	100%	-	-
German	TIGER	0.029	80%	**0.319**	80%
	Wikipedia	0.450	100%	**0.469**	100%
Spanish	AnCora	**0.299**	60%	**0.427**	80%
	Wikipedia	**0.445**	100%	**0.461**	100%
Romanian	Wikipedia	**0.610**	100%	**0.364**	100%

Table 5.4: The performance of the LDA models. Highlighted values indicate better performance than that of the frequency count baseline. Underlined values indicate better performance than that of the similarity-based method.

actions. The results confirm the previous results, with correlations above 0.550 for both Classical LDA and Rooth LDA. In the case of verb-subject pairs, which I showed in section 5.3.1 to be difficult to model, topic models performed better than both the corpus frequency baseline and the similiarity-based method across the board. In the case of verb-direct object interactions, the results obtained with Wikipedia as training corpora are comparable to, or better than the baselines, while among the two small training corpora, improvements over the baselines are only seen for AnCora. I attribute the poor results on the TIGER corpus to data scarcity, an explanations that was previously acknowledged in (Peirsman and Padó, 2010).

My results also confirm the claim made in (Ó Séaghdha, 2010) that Rooth LDA performs similarly, but slightly better overall than Classical LDA at modelling selectional preferences. As indicated in table 5.4, Spearman ρ

values for Rooth LDA are larger on almost all training corpus-test dataset combination. However, the correlation differences between the two LDA methods are small, at under 0.05 in all cases except on the Romanian verb-subject test dataset.

Addressing the coverage problem

While they perform very well at modelling selectional preferences, topic models have the major disadvantage of being unable to generalize over unknown verbs or arguments. Because setting the plausibility of verb-argument pairs containing unknown words to 0.00 makes it impossible to evaluate correlation with log-transformed test datasets, the two available alternatives are approximating the plausibility with a low, smoothing value or removing the tuples from the test data and reporting coverage along with correlation. The former method is mentioned in (Ó Séaghdha, 2010), while the latter was used in (Peirsman and Padó, 2010) with the similarity-based models.

To address this disadvantage, rather than assigning a default low plausibility to tuples containing unknown words, I proposed and tested a solution to replace the unknown words with similar words which are present in the training data. The idea of this technique is to use a word context vector space built with co-occurrence statistics from a very large POS-tagged corpus, and to search for similar words by identifying nearest neighbours under the cosine similarity. In this use case, POS-tagged data is a cheap resource to compute, given that a statistical POS tagger can easily be trained on data obtained from the dependency-parsed corpus for training the topic model.

In my implementation, I used the POS-tagged versions of the German and Spanish Wikipedias to build vector spaces containing all nouns and verbs in the two languages, and a context window of four words to either side as a co-occurrence statistic, where the context windows were not allowed to extend beyond sentence boundaries. For the dimensions of the vector space, I used the projection onto German and Spanish, respectively, of the noun and verb translation tuples extracted from Wiktionary, described in section 4.4.2.

I tested the unknown word approximation method on the training corpora which didn't fully cover the test datasets for their respective languages: TIGER for German, and AnCora for Spanish. Table 5.5 gives the correlation differences between assigning a low, 10^{-6} smoothing value to tuples containing unknown words, and substituting unknown words for known ones using the replacement algorithm. All results in the table are based on the Classical LDA variant of topic model.

Corpus	Algorithm	Direct Object		Subject	
		ρ	Coverage	ρ	Coverage
TIGER	Default 0.1	0.064	80%	0.212	80%
	Unknown Word Repl.	-0.126	100%	0.091	100%
AnCora	Default 0.1	0.128	60%	0.350	80%
	Unknown Word Repl.	**0.262**	100%	**0.411**	100%

Table 5.5: The comparative performance of defaulting the plausibility of tuples with unknown words to 10^{-6} and that of replacing unknown words through vector space similarity, tested with Classical LDA.

While the unknown word replacement method produced no improvements for German, I observed a significant improvement in performance for Spanish. This discrepancy can be explained when debugging the unknown words in the German dataset. Table 5.6 gives the list of unknown words, along with the replacement candidates found with the vector space. While the noun replacements candidates are, with two exceptions, similar terms from the point of view of selectional preferences, all of the verb replacement candidates are unrelated to the original verb. On further investigation, I concluded that the German vector space does not capture verb similarity well because the word order in the German language is relatively free, making it impossible for the algorithm to approximate syntactic similarity with word window contexts.

German

Direct Object Dataset				
Part of Speech	Unknown Word		Replacement	
Noun	Leichtigkeit	*(ease)*	Sicherheit	*(safety/certainty)*
	Rindfleisch	*(beef)*	**Unterschied**	*(difference)*
	Gehweg	*(footpath)*	Bürgersteig	*(footpath)*
Verb	**erlegen**	*(kill)*	**fressen**	*(eat)*
	schmieden	*(forge)*	**vereiteln**	*(frustrate)*
Subject Dataset				
Part of Speech	Unknown Word		Replacement	
Noun	Knabe	*(boy/lad)*	Junge	*(boy/youngster)*
	Grundschule	*(primary school)*	Bildung	*(education)*
	Bier	*(beer)*	Wein	*(wine)*
	Ferne	*(distance)*	Nähe	*(proximity)*
	Reifeprüfung	*(A-levels)*	**Film**	*(film)*
Verb	**knurren**	*(growl)*	**gehen**	*(go/walk)*
	glitzern	*(glitter)*	**übrigbleiben**	*(remain)*

Table 5.6: The unknown words in the German test dataset, along with their replacement candidates. Highlighted entries are mistakes made by the replacement algorithm.

5.4 Cross-lingual results

5.4.1 Baselines

I established the baselines for the performance of cross-lingual selectional preference transfer in the case of German, Spanish and Romanian by using the algorithms described in section 3.2.1. While I consulted native speakers of the three languages in order to increase the quality of the dictionary translations of the verbs and arguments in the test datasets, it should be noted that one-word translations were not possible for some of the terms, and that less specific translations were used in those cases: for example, the Spanish word *'tenista'*, meaning female tennis player, was approximated in English by the more generic term *'player'*.

All cross-lingual transfers were done with English as a target language, where

English plausibilities of the verb-argument pairs were estimated using Classical LDA trained on the BNC corpus. The baseline results are summarized in table 5.7, below.

Source Language	Direct Object		Subject	
	ρ	Coverage	ρ	Coverage
German	0.180	100%	0.052	100%
Spanish	0.548	100%	0.266	100%
Romanian	0.528	100%	0.313	100%

Table 5.7: The cross-lingual baseline for transferring selectional preferences from English using translations based on dictionary resources.

As the results in the table indicate, performance was below that of both Classical LDA and Rooth LDA when trained on the large dependency-parsed corpora (the parsed versions of Wikipedia for each language). However, when comparing to the smaller TIGER and AnCora, the cross-lingual transfer method outperforms the monolingual topic models on German *direct objects*, and Spanish *subjects*. This confirms that the cross-lingual methods have applicability for resource-poor languages, where either not enough parsed data is available to build high-quality selectional preference models, or not parsed data is available altogether.

5.4.2 Bilingual vector spaces

I built the bilingual vector spaces between English and the other languages starting from the instructions given in (Peirsman and Padó, 2010), and using the Wikipedias in the four languages as POS-tagged corpora. The vector space representation was derived using corpus-derived raw frequency counts over an empirically-determined context window of four tokens as the co-occurrence statistic. Word context windows are used because positional information captures association well, and provides the simplest way of approximating the syntactic context of a word in the absence of parsed data. For example, the subject of a verb is usually the noun immediately preceding

it, while the direct object of a verb is usually the noun following it. Although these heuristics have very low accuracy in languages with relatively free word order, such as German, they are still useful for inferring information about word similarity from an unparsed corpus. Two words represented in the resulting vector space are inferred to be a translation pair when they are included in the top k-nearest neigbours of each other, where only words belonging to the same part of speech, and the opposite language are included in the search. The original paper recommends setting $k = 1$, but I got the best results when setting $k = 3$.

During the development stage, I also tested pointwise mutual information and log-transformed pointwise mutual information as co-occurrence statistics, but I found the resulting distances in the vector space to be less indicative of lexical similarity than those obtained with raw frequency counts. Similarly, I experimented with different word context windows. In all cases, the context window was not allowed to extend beyond the edge of the sentences. While one and two-word context windows lead to vector spaces dominated mostly by noise, further increasing the context window beyond four words brings a degradation in performance due to the homogenization of the co-occurrence counts between the nouns and verbs. Although I only used symmetric word windows in my experiments, introducing pairs of left-spanning and right-spanning counts as a co-occurrence statistic is a direction of future research.

Unlike Peirsman and Padó, who used all of the open-class parts of speech in building the individual vector spaces, I only relied on verbs and nouns. The reasons for my choice are twofold. Firstly, while adjectives and adverbs can contribute to the similarity metrics, they almost double the size of the vector space, driving the time complexity of nearest-neighbour search up by a factor of 4, and making it impractical to keep the whole structure in memory. Secondly, it was considerably harder to obtain adjective and adverb seed translation pairs for labelling the axes of the vector spaces than it was for nouns and verbs.

5.4.3 Observations and improvements

I couldn't use my implementation to exactly replicate the results on cross-lingual selectional preferences as reported in (Peirsman and Padó, 2010), because the models used for predicting plausibilities in the target language are different. However, while trying to improve on the construction of the bilingual vector spaces, I identified three key aspects which were not mentioned in (Peirsman and Padó, 2010), but which are essential to building a bilingual vector space with good first-translation accuracy.

The first observation is that even though (Peirsman and Padó, 2010) claims that bilingual vector models can be built from unrelated corpora, I discovered that the first-neighbour translation accuracy of the models is greatly increased by using related corpora. The explanation for this observation is that in similar corpora, words which are translations of each other are guaranteed to share similar contexts in their respective languages, so that they can be correctly identified through vector distances. Thus, using Wikipedia is particularly advantageous when building bilingual vector spaces.

Besides the co-occurrence statistic used, the most important factor impacting the performance of bilingual vector spaces is the set of bilingual axes. The original authors of (Peirsman and Padó, 2010) suggest an inexpensive way of bootstrapping an informative set of axes by starting out with a set of word cognates compiled based on the edit distance, and then iterating until convergence. New translation pair candidates would be added to the initial dimension set after each iteration, while wrong translations in the seed set would be corrected. However, this method of building the vector spaces is not robust, because if inaccurate translation pairs are added by the bootstrapping early on, errors then quickly cascade and the resulting first-neighbour translations observed as early as the following iteration are of low quality. While the bootstrapping method can work when using the right translation seeds, in order to avoid problems caused by its brittleness, I bypassed the bootstrapping procedure altogether, and used the verb-noun balanced translation pairs extracted from Wiktionary as final dimension sets.

In order to address the brittleness issue, two pre-requisites must be met: there have to be enough translation seeds to exceed the *critical mass* needed to start and drive the bootstrapping process, and the composition of the translation seed set has to be balanced between the POS-tags considered, which in my implementation were verbs and nouns. While the first pre-requisite is intuitive, it is less intuitive that the seed translations should be balanced between verbs and nouns. According to my observations, verb dimensions provide more discriminative power for nouns than noun dimensions. This observation is supported by the fact that the vector space is built on context co-occurrences which are meant to approximate verb-argument interactions. The rule is more intuitive when phrased from the point of view of verbs: noun dimensions provide more discriminative power for them than verb dimensions, simply because verb-verb interactions in the corpus are rare, and mostly attributable to noise. This fact was already acknowledged in (Peirsman and Padó, 2010), where it is mentioned that "the context of verbs is dominated by their arguments". Thus, an informative bilingual vector space requires a POS-balanced dimension set.

Source Language	Direct Object		Subject	
	ρ	Coverage	ρ	Coverage
German	0.129	100%	**0.535**	100%
Spanish	0.152	100%	**0.481**	100%
Romanian	0.267	100%	0.107	100%

Table 5.8: The performance of the cross-lingual selectional preference from English. The English plausibilities were estimated with a Classical LDA model, trained on the BNC.

As summarized in table 5.8, vector space-driven cross-lingual transfer performed better on the German and Spanish subjects than both the cross-lingual baseline, and the native Wikipedia-trained language topic models. While correlation was lower for other test cases, the results were still significant. This confirms that cross-lingual plausibility transfer can outperform monolingual methods in cases where dependency-parsed training data is scarce.

5.5 The threshold for better performance with cross-lingual transfer

With some of the cross-lingual performance results reported in sections 5.4.1 and 5.4.2 comparing promisingly against the monolingual baselines, the final question that I wanted to address in my research is how the ranking of the studied methods changes when varying the amount of the dependency-parsed data available for training. It is important to determine the thresholds where the performance of cross-lingual transfer methods is overtaken by that of the monolingual in order to determine the applicability of cross-lingual methods to computing selectional preference plausibilities when working with resource-poor languages.

5.5.1 Experimental setup

In order to measure performance in a consistent fashion across all amounts of training data, I used the four languages' full test datasets in all measurements, regardless of the coverage. Because of this, the correlation measurements for the *small training data sizes* are noisy.

I controlled the amount of training data in my experiments by sampling sub-sets of the verb-argument co-occurrence matrices extracted from the dependency-parsed German, Spanish and Romanian Wikipedia. For a given verb-argument pair, I used a sampling probability proportional to the frequency of that pair in the original corpus. I selected the measurement points for the training data size at intervals of $2,500$ training instances between $2,500$ instances and $20,000$ instances, at $5,000$ instances between $20,000$ and $100,000$ instances, and at $100,000$ between $100,000$ and $500,000$ instances. Additionally, I also measured performance at $1,000,000$ training instances and beyond, at intervals of $500,000$ instances, until reaching the maximum amount of parsed data available. While for English, Spanish and German, I only plotted measurements up to $1,000,000$ instances, for readability and

uniformity purposes, I verified that the trends outside the plotted data are consistent with those below $1,000,000$ training instances. For ease of comparison and readability purposes, all charts plotted in figures 5.2 to 5.5 use a y-axis scale between -0.1 and $+0.7$.

5.5.2 Methods plotted

For English, the four methods plotted are the four monolingual baselines: the **Corpus frequency** baseline, estimating plausibilities from the corpus frequency counts, the **Similarity method** described in section 3.1.2, which uses vector distances in a syntactic vector space to compute the similarity function and corpus frequency counts to compute interpolation weights, and the two version of Latent Dirichlet Allocation, **Classical LDA** and **Rooth LDA**, described in sections 3.1.3 and 3.1.3, respectively.

For the three other languages, the charts also include the performance thresholds for the two cross-lingual selectional preference transfer methods using English as a target language: the **Dictionary translations** method, in which I used dictionary resources to obtain translations, and the **Bilingual vector spaces** method, in which the translations were obtained automatically. The two algorithms were described in detail in sections 3.2.1 and 3.2.2, respectively.

5.5.3 Measurements

For **English**, I plotted the performances of the four studied monolingual methods on direct objects and subjects in figure 5.2. Both the *direct object*, and the *subject* chart confirm my observation that the performance of similarity methods does not improve substantially with increasing the amount of training data beyond $50,000$ verb-argument instances, while direct corpus frequency estimations improve monotonously with increasing amounts of training data. The charts also confirm that the two version of LDA implemented are almost identical in terms of performance, and that $200,000$

training instances is the limit beyond which increasing the amount of data does not result in performance gains.

In the case of **German**, the measurements for the two grammatical relations are plotted in figure 5.3. The German plots indicate that for amount of training data under 50,000 verb-argument instances, the LDA correlation measurements are noisy due to reduced coverage. This observation fully explains my inability to obtain good correlations when training on the TIGER corpus, which falls in the below-50,000 instances category. The relative ranking of the monolingual methods is similar to the English one, with the notable exception that corpus frequencies perform worse than the similarity model on *direct objects*, perhaps because of the noise in the automatically-parsed corpus and the increased lexical diversity of German brought by compound nouns. In the case of *direct objects*, we start obtaining better predictions with monolingual models than with bilingual vector spaces when training on more than 100,000 instances, while in the case of *subjects*, the performance of bilingual vector spaces is not reached by monolingual models before ceiling out.

The performance characteristics on **Spanish**, plotted in figure 5.4, follow the same patterns as the German ones: topic models are noisy on amounts of training data below 50,000 instances (but only in the case of subjects), and topic models rank higher than the other monolingual methods with large amounts of training data. Like German, bilingual vector spaces are not overtaken by monolingual models in the case of *subjects*, but are rapidly overtaken by all monolingual models on *direct objects*.

The measurements for **Romanian**, plotted in figure 5.5, are especially interesting from the point of view of a resource-poor language. The same observations as in the case of German and Spanish, regarding noise with low-data measurements, also apply for Romanian. I only had enough dependency-parsed data for 500,000 training instances, so I could not determine whether LDA models overtake corpus frequency counts at 1,000,000 training instances, although the trends in the charts suggest it. In terms of cross-lingual transfer, just as in the case of Spanish and German, bilingual vector spaces

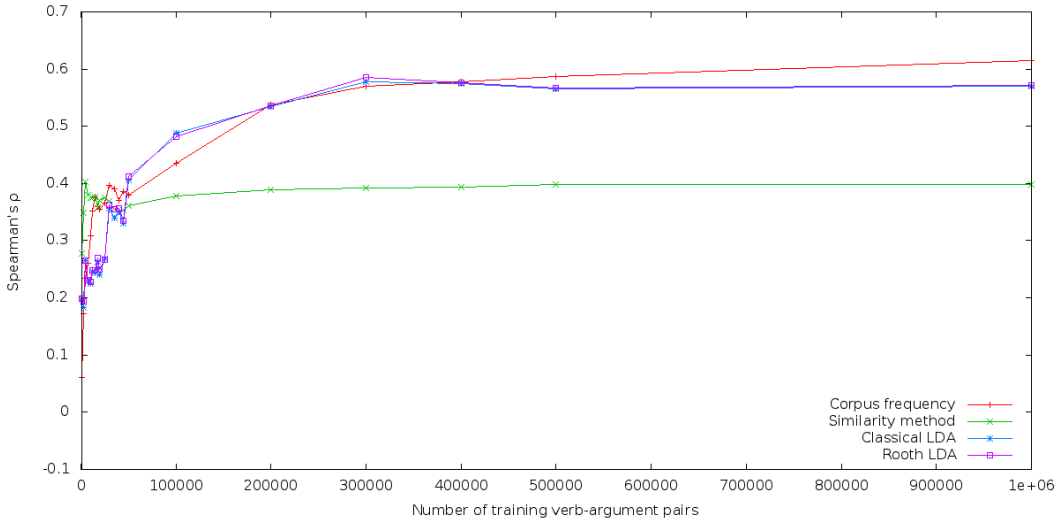
perform worse than all monolingual models on direct objects, but competitively on subjects.

5.5.4 Conclusions

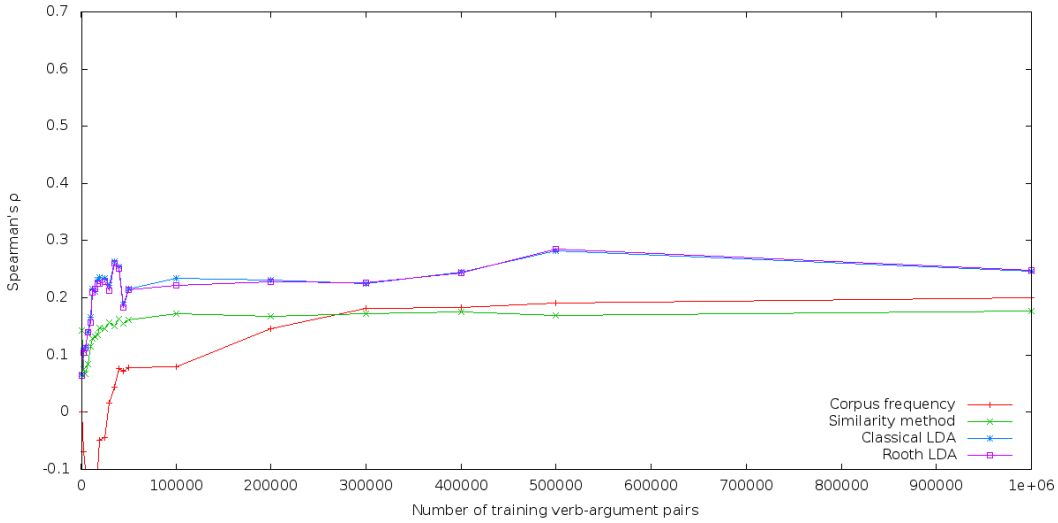
Following the experiments described in this chapter, a few conclusions can be drawn regarding the robustness of the various models of selectional preference, as well as regarding the overall performance of the cross-lingual plausibility transfer methods.

With regards to the **monolingual methods**, I observed that the two generative models used, Classical LDA and Rooth LDA, are unreliable and perform worse than the studied similarity based method and corpus frequency method when trained on a small number of instances. The minimum limit beyond which they show poor performance was highest for German, at $50,000$ training instances. However, when trained on large amounts of data, the generative models achieve better correlations than the other two models. A possible explanation for the bad performance on small datasets could be the need to use a smaller number of classes when working with little data. On the other hand, I have shown that in most cases, the similarity-based model performs better than the other monolingual models with small amounts of training data, but does not improve considerably when increasing the amount of training data.

With regards to the **cross-lingual methods**, I observed that better correlations are achieved with automatic translations than with dictionary translations in the case of *subject* relations (except on the Romanian data), and that this pattern is reversed for *direct object* relations. The experiments also showed that cross-lingual transfer performs better than, or similar to the monolingual models on *subject* relations, and while it is worse than the monolingual models on *direct object* relations, it is still useful in cases where no dependency-parsed data is available.

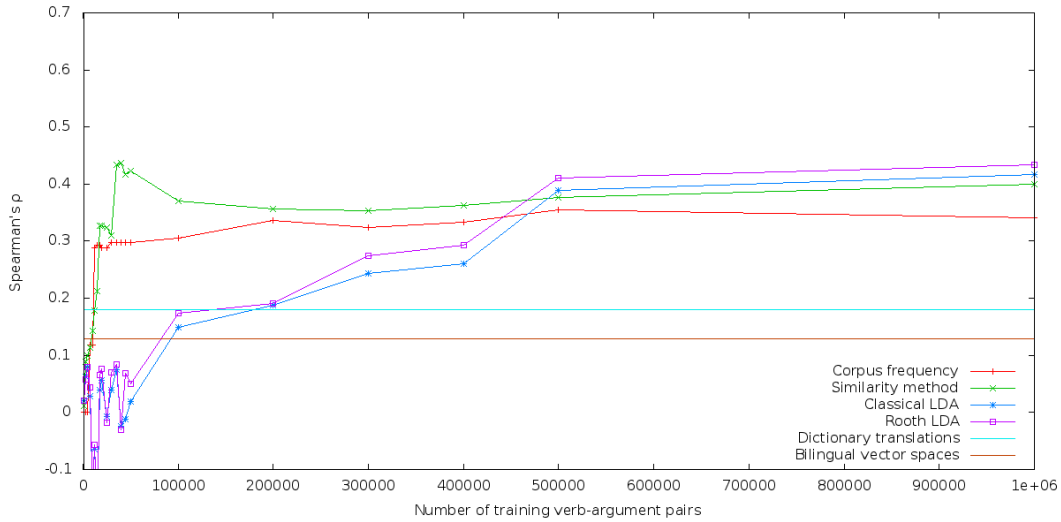


(a) Direct object

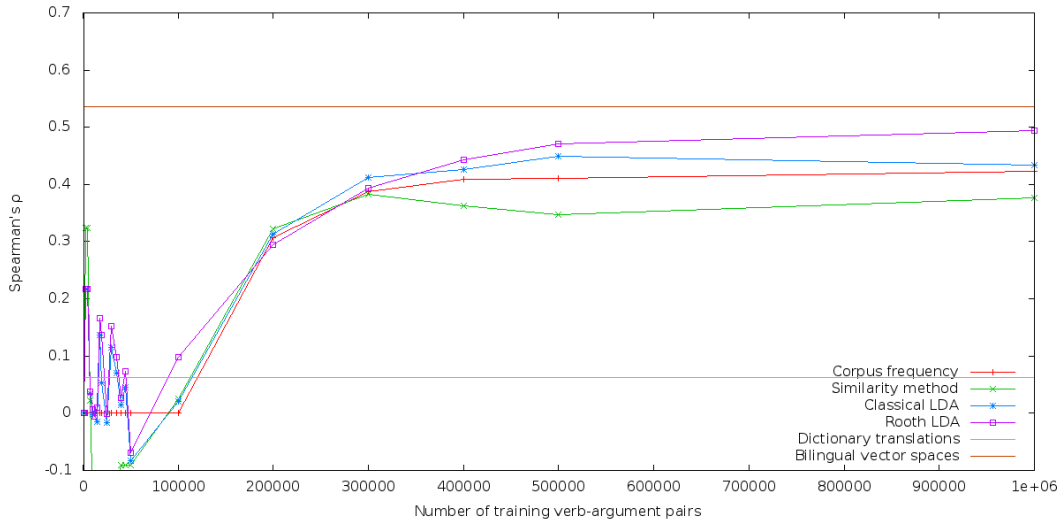


(b) Subject

Figure 5.2: The performance of English monolingual selectional preference models, plotted against the volume of data used to train them.

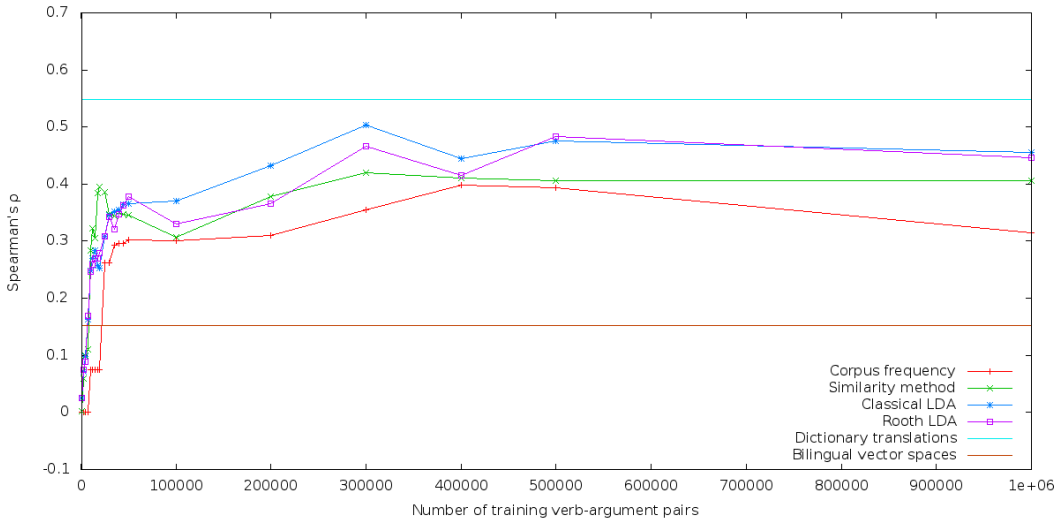


(a) Direct object

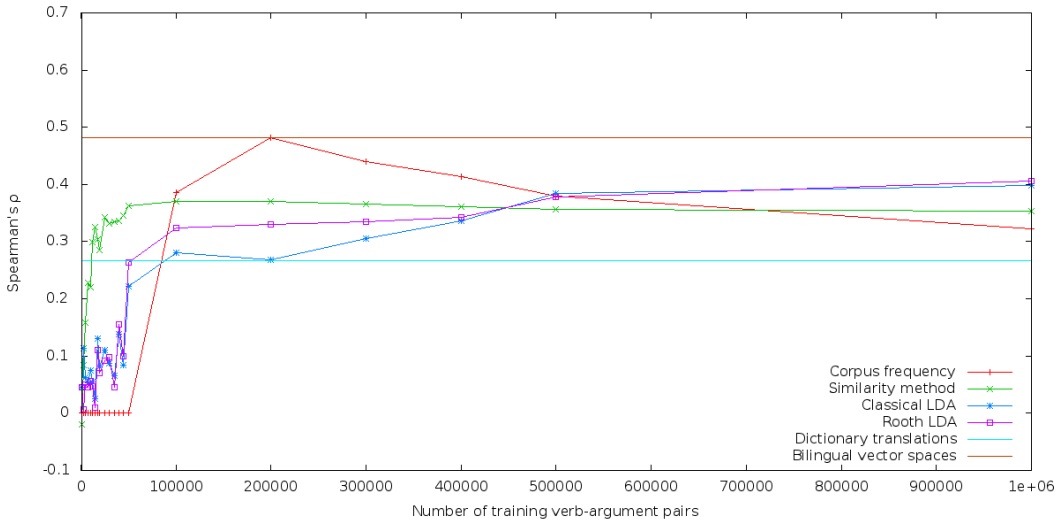


(b) Subject

Figure 5.3: The performance of German monolingual selectional preference models, plotted against the volume of data used to train them. Cross-lingual transfer performances based on a BNC-trained, English Classical LDA model are included as thresholds, for reference.

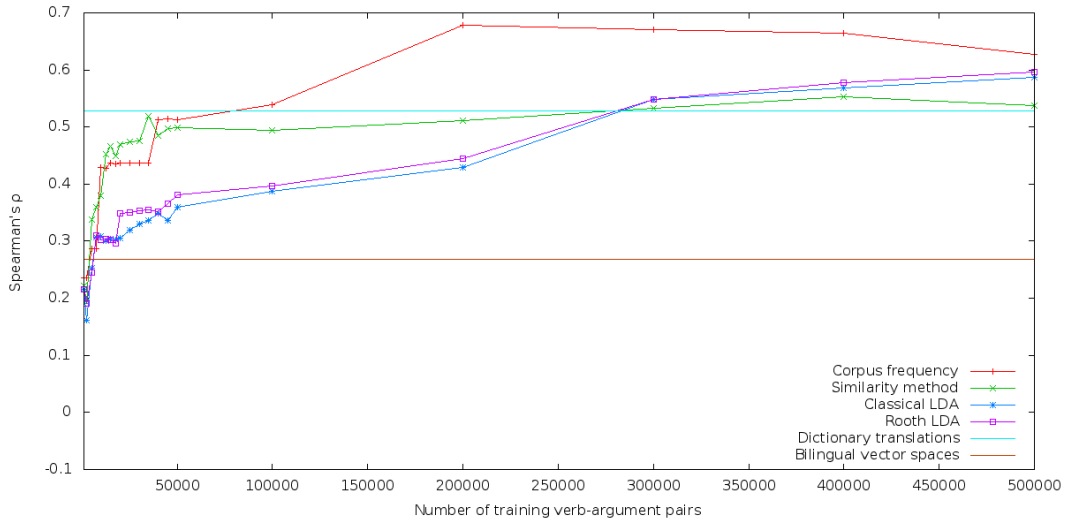


(a) Direct object

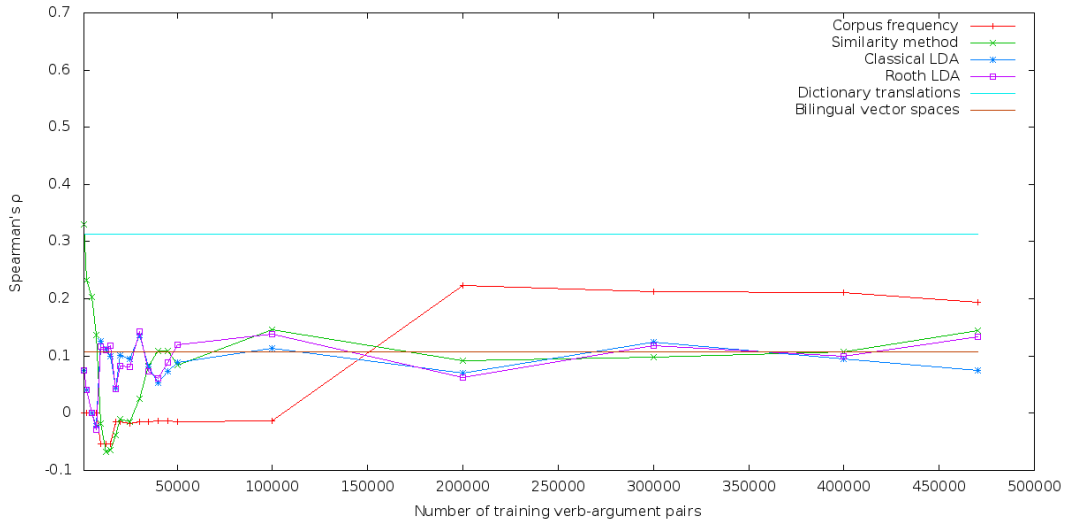


(b) Subject

Figure 5.4: The performance of Spanish monolingual selectional preference models, plotted against the volume of data used to train them. Cross-lingual transfer performances based on a BNC-trained, English Classical LDA model are included as thresholds, for reference.



(a) Direct object



(b) Subject

Figure 5.5: The performance of Romanian monolingual selectional preference models, plotted against the volume of data used to train them. Cross-lingual transfer performances based on a BNC-trained, English Classical LDA model are included as thresholds, for reference.

5.6 Error analysis and discussion of limitations

In order to better understand the issues affecting the performance of the cross-lingual plausibility transfer methods, I performed an error analysis and identified four main limitations which I am describing in this section.

My research hypothesis claiming that the cross-lingual transfer of selectional preferences can benefit resource-poor languages is based on the observation that selectional preferences are *largely* language-independent logical consequences of world facts. However, they are not *entirely* language-independent. A linguistically obvious limitation to the method's success are idioms, which do not carry across languages. For example, the four transitive expressions listed in table 5.9 are all idioms for the expression *to die*, and constitute plausible verb-direct object combinations in their respective languages. Their English word-for-word translations, however, are not equally likely combinations.

Language	Idiom	English word-for-word translation
English	**to kick** *the bucket*	-
German	*den Löffel* **abgeben**	**to give away** *the spoon*
Spanish	**estirar** *la pata*	**to stretch** *the foot/leg*
Romanian	**a da** *ortul*	**to give/pay** *the penny*

* verbs are in bold, while direct objects are in italic font.

Table 5.9: Example of four different idioms containing transitive verbs for the expression *to die*, along with their English word-for-word translations.

A similar, but easier to address limitation is the independent translation of the verb and its argument into another language where more than one translation is possible for either of the words. This problem became immediately visible when compiling the translations for the cross-lingual baseline using dictionary resources. For example, the pair '*Subvention* **kürzen**', taken from the German *direct object* gold standard, was initially translated to '**to shorten** *subvention*', while a more accurate, argument-aware translation of

the verb would have been '**to cut**'. Unlike idioms, this problem can be overcome by translating the verb and its argument together.

A more serious issue limitation of the model coming from linguistic differences between the source and the target language is the impossibility to accurately capture the meaning of a word with a single-word translation. This problem is very pronounced between the lexically-agglutinative German and the lexically-isolating English. For example, no single word in English can accurately capture the meaning of German *'Kochbuch'* (EN: *'cook book'*), or *'Gesellschaftswissenschaft'* (EN: *'social science'*). However, it is present to some extent between all languages. For example, there is no single-word English translation for Spanish *'tenista'* (EN: *'female tennis player'*).

The final limitation I observed, and which was also mentioned in (Peirsman and Padó, 2010), is that translations from bilingual vector spaces built on context co-occurrences tend to provide translations which are **semantically related**, rather than **semantically similar** to the original word. For example, the noun *'photographer'* is semantically related to the noun *'wedding'*, but the two are not similar because they cannot typically fill the same arguments slots for verbs; an example of a noun similar to *'photographer'* is *'artist'*. This limitation could be lessened by learning to predict the position of arguments relative to the verb, in order to better approximate a syntactic context vector space.

5.7 Summary

In the first part of this chapter, I described the testing methodology I used for evaluating the performance of the models studied, I described the existing test datasets for English, German, and Spanish, and I showed the steps I took for compiling and releasing the test dataset for Romanian.

In the second part of the chapter, I described the implementation details and the performance results of the monolingual and cross-lingual models studied, when trained on each of the available corpora for the four languages, and

showed that LDA-based methods are bestter than the other monolingual methods, but that similarity-based methods are robust to small amounts of training data. I also introduced a lexical substitution method based on syntactic vector spaces to improve LDA coverage.

In the third part of the chapter, I described the experiment I designed for studying how well the cross-lingual models compare to monolingual models depending on the amount of dependency-parsed data available, and presented the conclusions of the measurements, showing that cross-lingual transfer outperforms monolingual models on subjects and is informative on direct objects. In the final part of the chapter, I presented the conclusions and limitations identified during error analysis.

Chapter 6

Conclusions

6.1 Summary

This project aimed to verify the extent to which transferring verb-direct object and verb-subject selectional preference plausibilities across languages can outperform the estimations made in the native language under conditions of data scarcity, and to determine the applicability of the cross-lingual models to resource-poor languages where no monolingual alternatives exist.

As part of my monolingual investigation, besides English I studied the morphologically different German and Spanish languages, as well as the truly resource-poor Romanian language, and compiled and published a gold standard of human plausibility judgements for the latter. I described how Wikipedia can be used as a publicly-available data source based on which to compile automatically-parsed corpora which are on the scale of the English-language BNC, and used the resulting corpora to train and test two state of the art topic models of selectional preference which had been previouly only tested on English. I varied the amount of training data, and discovered that while topic models can achieve good performance, ranking above the other monolingual baselines in most experiments, they are unreliable when trained on less than $50,000$ verb-argument interactions. I observed this limitation for

all corpora, regardless of the language, and I proposed and tested a lexical substitution method which helped topic models deal with unknown words in the lexically-isolating Spanish, in conditions of training data sparsity.

In the second part of the project, I conducted cross-lingual transfer experiments using translations taken from dictionary resources as a baseline, and bilingual vector spaces as an automatic translation method, trainable with large amounts of unparsed data. I described a more computationally-effective, Wiktionary-based alternative to setting the dimensions for the vector space than the one described in the literature, and identified key factors influencing performance. I then experimentally confirmed that bilingual vector spaces outperform monolingual methods when modelling verb-subject interactions, as well as being effective when little or no parsed training data is available.

Last but not least, I performed an error analysis and identified four main limitations which need to be addressed by future models of cross-lingual selectional preference transfer to obtain improved performance.

6.2 Directions for future work

Due to the limited time scope of the project, not all of the possibilities for performance improvement could be investigated. While the results reported for cross-lingual selectional preference transfer are promising, continued research could bring further improvements. I identified four main directions of future work.

Firstly, as I indicated in the error analysis section, two of the main performance limitations could be overcome by taking into account both the verb and the argument when searching for translations, rather than translating them individually. Secondly, future work could measure the performance benefits derivable from including adjectives and adverbs in the construction of the bilingual vector spaces. Thirdly, investigations into the performance impact of using asymmetric word context windows could determine the ex-

tent to which syntactic context vector spaces can be approximated with word context vector spaces. Previous research done by Peirsman and Pado suggests learning the parameters of the word context window through semi-supervised learning from a small, dependency-parsed corpus, in addition to the unparsed data. Finally, a fourth suggested direction of further research would be to explore methods of computing plausibility estimations which involve more than one neasrest neighbour in the target language.

Bibliography

Shane Bergsma, Dekang Lin, and Randy Goebel. Discriminative learning of selectional preference from unlabeled text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 59–68, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

D.M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4): 77–84, 2012.

D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The tiger treebank. In *Proceedings of the workshop on treebanks and linguistic theories*, pages 24–41, 2002.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. Tiger: Linguistic interpretation of a german corpus. *Research on Language and Computation*, 2(4):597–620, 2004.

Ted Briscoe, John Carroll, and Rebecca Watson. The second release of the rasp system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 77–80. Association for Computational Linguistics, 2006.

Carsten Brockmann and Mirella Lapata. 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*, pages 27–34. Association for Computational Linguistics, 2003.

Nathanael Chambers and Dan Jurafsky. Improving the use of pseudo-words for evaluating selectional preferences. In *Proceedings of the 48th Annual*

Meeting of the Association for Computational Linguistics, pages 445–453. Association for Computational Linguistics, 2010.

Noam Chomsky. *Syntactic structures*. Mouton, 1957.

Stephen Clark and David Weir. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206, June 2002. ISSN 0891-2017.

Ido Dagan, Lillian Lee, and Fernando CN Pereira. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69, 1999.

Kathrin Erk. A simple, similarity-based model for selectional preferences. In *In Proceedings of ACL-07*, pages 216–223, 2007.

Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, September 2002. ISSN 0891-2017.

Ralph Grishman and John Sterling. Acquisition of selectional patterns. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 658–664, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. doi: 10.3115/992133.992172. URL `http://dx.doi.org/10.3115/992133.992172`.

Ralph Grishman and John Sterling. Smoothing of automatically generated selectional constraints. In *Proceedings of the workshop on Human Language Technology*, pages 254–259. Association for Computational Linguistics, 1993.

Samer Hassan and Rada Mihalcea. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1192–1201, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-63-3.

D. Hindle and M. Rooth. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120, 1993.

Frank Keller and Mirella Lapata. Using the web to obtain frequencies for unseen bigrams. *Computational linguistics*, 29(3):459–484, 2003.

E. Lefever and V. Hoste. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20. Association for Computational Linguistics, 2010.

Montserrat Marimon, Beatriz Fisas, Núria Bel, Blanca Arias, Silvia Vázquez, Jorge Vivaldi, Sergi Torner, Marta Villegas, and Mercè Lorente. The iula treebank. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC2012)*, 2012.

Maria Antònia Martí, Mariona Taulé, Manu Bertran, and Lluís Màrquez. Ancora: Multilingual and multilevel annotated corpora. *MS, Universitat de Barcelona*, 2007.

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. 2002.

Diana McCarthy and John Carroll. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654, December 2003. ISSN 0891-2017.

Ken McRae, Michael J Spivey-Knowlton, and Michael K Tanenhaus. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312, 1998.

R. Mihalcea, R. Sinha, and D. McCarthy. Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14. Association for Computational Linguistics, 2010.

R. Navigli and S.P. Ponzetto. Babelrelate! a joint multilingual approach to computing semantic relatedness. In *Proceedings of the 26th Conference on Artificial Intelligence (AAAI-12)*, 2012.

Joakim Nivre, Johan Hall, and Jens Nilsson. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219, 2006.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kubler, Svetoslav Marinov, and Erwin Marsi. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95, 2007.

Diarmuid Ó Séaghdha. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 435–444, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Sebastian Padó, Ulrike Padó, and Katrin Erk. Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of EMNLP/CoNLL*, volume 7, 2007.

Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

Yves Peirsman and Sebastian Padó. Semantic relations in bilingual lexicons. *ACM Trans. Speech Lang. Process.*, 8(2):3:1–3:21, December 2008. ISSN 1550-4875.

Yves Peirsman and Sebastian Padó. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 921–929, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5.

Philip Stuart Resnik. *Selection and information: a class-based approach to lexical relationships*. PhD thesis, Philadelphia, PA, USA, 1993. UMI Order No. GAX94-13894.

A. Ritter, O. Etzioni, et al. A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434. Association for Computational Linguistics, 2010.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 104–111, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics. ISBN 1-55860-609-3.

Beatrice Santorini. Part-of-speech tagging guidelines for the penn treebank project (3rd revision). 1990.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. Vorläufige guidelines für das tagging deutscher textcorpora mit stts. In *Draft. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung/Universität Tübingen, Seminar für Sprachwissenschaft*, 1995.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK, 1994.

Helmut Schmid. Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*. Citeseer, 1995.

D.O. Séaghdha and A. Korhonen. Probabilistic models of similarity in syntactic context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1047–1057. Association for Computational Linguistics, 2011.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. A new hybrid dependency parser for german. *Proc. of the German Society for Computational Linguistics and Language Technology*, pages 115–124, 2009.

Radu Simionescu. Hybrid pos tagger. *Language Resources and Tools with Industrial Applications*, page 21, 2011.

M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

Mariona Taulé, Maria Antonia Martı, and Marta Recasens. Ancora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*, 2008.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.