Leveraging Training Data from High-Resource Languages to Improve Dependency Parsing for Low-Resource Languages

Claire Jaja

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2014

Committee:

Fei Xia, Chair

Gina-Anne Levow

Program Authorized to Offer Degree:
Computational Linguistics

University of Washington

Abstract

Leveraging Training Data from High-Resource Languages to Improve Dependency Parsing for Low-Resource Languages

Claire Jaja

Chair of the Supervisory Committee:
Associate Professor Fei Xia
Linguistics

Dependency parsing is an important natural language processing (NLP) task with many downstream applications, and as is common in the field, high accuracy results can be obtained when using statistical methods and training on high-quality annotated training data. When dealing with low-resource languages where annotated training data is not available and prohibitively expensive to obtain, more clever methods must be used to leverage existing resources. My work in this thesis focuses on instance selection, which rests on the assumption, little explored cross-linguistically but well-proven monolingually in domain adaptation, that using less training data that is more relevant to your test case is better than using a full pool of potentially highly irrelevant training data. I conduct a larger, more thorough exploration than has previously been attempted into instance selection based on the perplexity of part-of-speech tag sequences, using the Google Universal Dependency Treebank, which spans ten languages. Additionally, I leverage another instance selection technique based on cross-entropy difference, which has shown superior results to perplexity selection when used for domain adaptation. These methods are both applied to two different potential pools of training data, one being the combination of multiple source languages, the other being English alone. Lastly, I explore automatic rearrangement of

the part-of-speech tags in the English training data to better match three potential target languages. These experiments show mixed results, which may help to inform future exploration in dependency parsing for low-resource languages. When a pool of multiple source languages is used, a significant boost is seen for target languages where relevant training data is available but infrequent in the training data, with cross-entropy difference providing slightly better performance than perplexity selection. However, these methods don't provide the same large improvements for target languages where lots of relevant training data is available among the multiple source languages or when English alone is used as the training data. Rearranging the part-of-speech tags has a small positive impact on the scores when using the entire training dataset, which is promising for more extensive rearrangement. However, applying instance selection methods to select training data from this rearranged data does not yield better results than selecting training data from the non-rearranged data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

Dependency parsing is a task where the input is a sentence and the output is a dependency tree. In the dependency tree, each lexical item has one (and only one) head that it depends on, with, optionally, a label describing the dependency relation. An example of a short sentence, taken from McDonald et al. (2005b), is shown in Figure 1.1, parsed using the Stanford dependency guidelines.

Dependency parsing is an important task in natural language processing (NLP), which feeds many downstream applications, including but not limited to relation extraction, question answering, machine translation, and semantic analysis. Often dependency parsing is preferable to traditional phrase structure parsing, since it provides information about the predicate-argument structure and is better able to handle free word order languages.

As with many tasks, current approaches to dependency parsing are dominated by statistical methods which rely heavily on large quantities of high-quality annotated training data. However, generating this annotated data requires time, money, and expertise, and as a result, building a dependency parser for a new language is often prohibitively expensive. Alternative methods of approaching this task for low-resource languages by leveraging annotated data from high-resource languages is a ripe area of research, and one which I explore in this thesis.

Previous work in this area is extensive, with a variety of techniques with varying success spanning more than a decade. These techniques range from the very simple approach of direct transfer, where dependency parsed source language data is delexi-

Input: John_1 hit_2 the_3 ball_4 with_5 the_6 bat_7

Output:

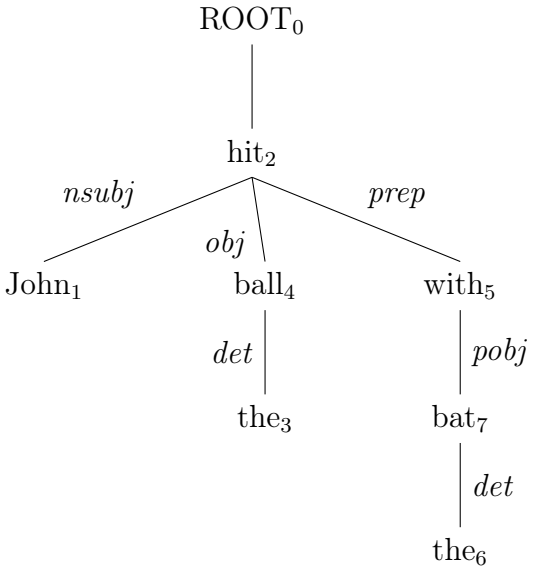

Figure 1.1: An example dependency parse

calized and used to train a parser which can be run on delexicalized target language data, to much more complicated approaches using projected transfer which rely on source-target parallel data so that dependency parses from the high-resource source language can be used to learn dependency parses for the parallel target language sentences. Recently, an approach inspired by domain adaptation has been attempted, with automatic instance selection being applied to delexicalized source language data. This approach is the one I build my own work on.

The approach explored in this thesis is a natural extension of previous work on applying instance selection techniques from domain adaptation. I evaluate the results of using part-of-speech (POS) tag perplexity to select training data instances on a novel dataset, the Google Universal Dependency Treebank, which spans ten languages. Additionally, I leverage another instance selection technique based on cross-entropy difference, which has shown superior results to perplexity selection when used for

machine translation domain adaptation. These methods are both applied to two different potential pools of training data, one being the combination of multiple source languages, the other being English alone. Lastly, I explore automatic rearrangement of the POS tags in the English training data to better match three potential target languages.

These experiments show mixed results, which may help to inform future exploration in dependency parsing for low-resource languages. When a pool of multiple source languages is used, a significant boost is seen for target languages where relevant training data is available but infrequent in the training data, with cross-entropy difference providing slightly better performance than perplexity selection. However, these methods don't provide the same large improvements for target languages where lots of relevant training data is available among the multiple source languages or when English alone is used as the training data. Rearranging the POS tags has a small positive impact, which is promising for more extensive rearrangement. However, applying instance selection methods to select training data from this rearranged data does not yield better results than selecting training data from the non-rearranged data.

Chapter 2
LITERATURE REVIEW

Previous work on dependency parsing for low-resource languages has included a wide variety of techniques. The simplest approach of direct transfer relies on information available in POS tags that may be relevant across languages. Most work, however, has centered on different methods of projected transfer, which leverage unparsed source-target parallel data to learn something about the dependency structure of the target language based on projecting parses from the parallel source language. This approach, however, hinges on the availability of source-target parallel data, as well as high-quality word alignments. A less explored but promising approach explores automatic selection of relevant training instances from a pool of delexicalized training data, inspired by the success of such methods in the context of domain adaptation.

2.1 Direct Transfer

The main idea of direct transfer is quite simple; POS tags hold significant information that is relevant cross-linguistically. In order to leverage this, lexical features are removed from source language dependency parsed data, and then a dependency parser is trained on this delexicalized data. This "language agnostic" dependency parser can then be tested on a similarly delexicalized target language. Results from using this technique are demonstrated in Zeman and Resnik (2008), McDonald et al. (2011), and McDonald et al. (2013).

The advantages of this technique are that it doesn't require any data from the target language at train time, and that the implementation is very simple. Additionally, results are quite good on closely related languages, as McDonald et al. (2013) show,

with unlabeled attachment scores for Germanic languages when trained on another Germanic language within 10 - 20% of training on the target language itself and for Romance languages when trained on another Romance language, within 6% of training on the target language itself. However, the performance degrades substantially with less related languages, as is the case in McDonald et al. (2013) when training on Romance or Germanic languages and testing on Korean, with unlabeled attachment scores 30 - 40% lower than training on Korean itself. Another hurdle is that the target language data needs to be POS tagged for testing; often results are reported using gold standard POS tags for the test data, but in real usage, the POS tagging would need to be done using automatic methods. This has the potential to introduce more errors, especially when testing on low-resource languages that may not have reliable POS taggers available.

2.2 Projected Transfer

Often, other resources may be available for the target language which can be leveraged for this task. Specifically, unparsed source-target parallel data can be used to learn the dependency structure of the target language. To do this, Hwa et al. (2004) dependency parse the source language side of the parallel data, then word align the parsed data with the target language side. The source language parse is then projected onto the target language side, a small set of handwritten language-specific post-project transformation rules are applied, and the resulting parses are filtered to eliminate likely to be bad trees, then used to train a target language dependency parser. This method yields results for Spanish that are comparable with a state of the art rule-based commercial system, with a F-score of 72.1%, while results are not as promising for Chinese, with a F-score of 53.9% achieved.

Alternatively, different methods of projecting the parses can be used. Smith and Eisner (2009) use unsupervised quasi-synchronous grammar projection, which improves on parsers trained using only high-precision projected annotations and far

outperforms, by more than 35% absolute dependency accuracy, unsupervised Expectation Maximization (EM), and has the advantage of not requiring handwritten language-specific transformation rules. Ganchev et al. (2009) use linear expectation constraints to allow partial, approximate transfer of parses, reaching unlabeled attachment scores of 78.3% for Bulgarian and 72.8% for Spanish when projecting from English and using a very small number of transformation rules (7 for Bulgarian, 3 for Spanish) to cover the most common structural differences. McDonald et al. (2011) predict target language parses using a direct transfer parser, then select the parses that best align with the source language parses to iteratively seed a new parser, with an average boost of 2.7% over direct transfer for 8 different target languages.

Overall, this method has a substantial performance boost over the direct transfer technique, and there are lots of clever methods for using the projected parses. Additionally, it doesn't (necessarily) require POS tagging the target language test data. However, this hinges on the availability of source-target parallel data. The word alignment step between the source and target sides of the parallel data can also introduce errors, and not all words can be aligned one to one - some are aligned one to many, many to one, or not aligned to any word whatsoever on the other side (as is often the case with grammatical function words).

2.3 Instance Selection

An interesting new approach has emerged recently, although it is little studied in this context. Instance selection, a common strategy within domain adaptation, rests on the idea that you can use less data if it's better data. To do this, the source language training instances that are closest to the target language are selected and used for training.

Yasuda et al. (2008) introduce the use of perplexity based instance selection for domain adaptation in machine translation, where they train a language model on either the source or target side of an in-domain corpus, then use the model to calculate

perplexity for each sentence from an out of domain parallel corpus and select as training data those with perplexity below a certain threshold. They are able to reduce the translation model size by 50% using this method, while improving the BLEU score by 1.76%. Moore and Lewis (2010) introduce instance selection using cross-entropy difference, where the difference between the cross-entropy on an in-domain language model and that on an out-of-domain language model is taken, to build domain-specific language models, and Axelrod et al. (2011) extend this method to machine translation, where selecting training data using this method results in translation models that consistently outperform general-domain baseline while using as few as 35k out of 12 million sentences, with an increase of 1.8 BLEU points.

This approach doesn't require parallel data, but it's not immediately clear what the best measure for determining similarity between the training and test data is when they come from different languages, since the selection techniques used in domain adaptation are based on words. However, Søgaard (2011) shows promising results using a perplexity based instance selection technique on four very different languages - Arabic, Bulgarian, Danish, and Portuguese. He trains a language model using POS tag sequences from the target language and uses the perplexity per word of the POS tag sequences from the remaining three source languages to determine the sentences that are most similar to the target language, then uses the 90% most similar sentences to train a dependency parser. The results vary by language, with no improvement over the baseline (training on the full delexicalized training data from the three source languages) for Danish, a small improvement for Arabic (from 45.5% to 48.4%), and a large improvement for Bulgarian (from 44.5% to 70.2%) and Portuguese (from 37.1% to 75.1%).

Techniques for dependency parsing for low-resource languages have varied from simple direct transfer relying on POS tags applying cross-linguistically to complex strategies for projecting and transforming parses based on parallel data. The promising, but under-explored, technique of instance selection has shown very large improve-

ments over direct transfer, while maintaining a fairly simple and easily extensible workflow. This warrants further investigation of adapting instance selection methods from domain adaptation, where it's been successful in language modeling and machine translation, to selecting high-resource source language sentences for dependency parsing a low-resource target language.

Chapter 3

METHODOLOGY

The approach I take in this thesis builds on previous work applying instance selection techniques from domain adaptation to the problem of dependency parsing for low-resource languages. My hypothesis is that clever methods of selecting source language training data instances will lead to better dependency parses for low-resource target languages by minimizing irrelevant and potentially misleading data in the training dataset. I leverage a novel dataset, the Google Universal Dependency Treebank, which spans ten languages and uses homogeneous syntactic dependency annotation. From this dataset, I define two different potential pools of training data for a given target language. One pool is a combination of the training data for every other available language; this allows for relevant training instances to be found across a variety of languages. The other pool is the full English training data, which is by far the largest training dataset and allows for results that may transfer to work on solutions for low-resource languages in a variety of tasks, since English training data is highly prevalent and likely to be available for many tasks.

First, I conduct a larger, more thorough exploration than has previously been attempted into instance selection based on the perplexity of POS tag sequences. Additionally, I leverage another technique for instance selection based on cross-entropy difference which has shown superior results to perplexity selection when used in the realm of domain adaptation. This work is the first to apply this method cross-linguistically. Lastly, I automatically rearrange POS tags in the English training data to better match three potential target languages. This rearrangement is limited to modifying adjectives, which are rearranged from preceding the noun they modify,

as is typically the case in English, to following the noun they modify, which more closely matches Indonesian, Spanish, and French. Both instance selection methods are applied to the rearranged training data.

3.1 Data

I use the Google Universal Dependency Treebank v2.0 (v1.0 described in McDonald et al. (2013)). This collection includes homogeneous syntactic dependency annotation for eleven languages (Brazilian Portuguese, English, Finnish, French, German, Italian, Indonesian, Japanese, Korean, Spanish, and Swedish), which allows for easy comparison of results on different languages. Two sets of POS tags are given for every language; one set uses the universal POS tags, as described in Petrov et al. (2012), while the other gives more fine-grained language-specific POS tags. The annotation used is a modified version of Stanford style dependencies with changes made to accommodate all the included languages; for a complete description of these changes, refer to McDonald et al. (2013). It is also freely available online.[1]

For my experiments, I use the "standard" version of the data, which does not include Finnish. This version matches most closely to standard Stanford dependencies; the representations are mostly content-head except that adpositions are the head in adpositional phrases and copular verbs are the head in copular constructions. I discard the fine-grained POS tags and only use the universal tagset. This data also comes with a standard split into training, development, and test sets. I only leverage the training and test sets in my experiments.

These ten languages span multiple language families and include three Germanic languages (German, English, and Swedish), four Romance languages (Spanish, French, Italian, and Brazilian Portuguese), and three non-Indo-European languages (Indonesian, Japanese, and Korean). This wide coverage allows me to test how robust my

[1] https://code.google.com/p/uni-dep-tb/

methods are; some techniques in previous work were only tested on highly related languages (e.g. only Romance languages) and thus may not work well on less related languages. See Figure 3.1 for a diagram of the linguistic relation between the languages. From here out, ISO 639-1 codes are used for the ten languages in all figures and tables.

Some statistics about the datasets are shown in Table 3.1. In particular, note that the training and test sets vary in size, as well as in average sentence length.

Most notably, English has the largest quantity of training data with nearly three times as many sentences as the next largest training set. German, Spanish, and French all have comparably sized training sets. By comparison, the training sets for Swedish, Italian, Brazilian Portuguese, Indonesian, Japanese, and Korean are quite small with none of them exceeding 10,000 sentences; they are all less than a quarter the size of the English training set.

Also of interest is the variance in average sentence length. Japanese and Korean both have exceptionally short sentences, averaging around 10 tokens per sentence. This is likely due to some combination of the nature of the languages themselves, the tokenization schema used, and the peculiarities of this particular dataset. Average sentence length in all but two other languages exceeds 20 tokens per sentence. German and Swedish fall between the two extremes with average sentence lengths around 16 tokens per sentence.

3.2 Parser

I use the MSTParser v0.5.1 described in McDonald et al. (2005a), McDonald et al. (2005b), McDonald and Pereira (2006), and McDonald et al. (2006) and freely available online.[2,3] This parser is a non-projective dependency parser that searches for

[2]http://www.seas.upenn.edu/~strctlrn/MSTParser/MSTParser.html

[3]http://sourceforge.net/projects/mstparser/

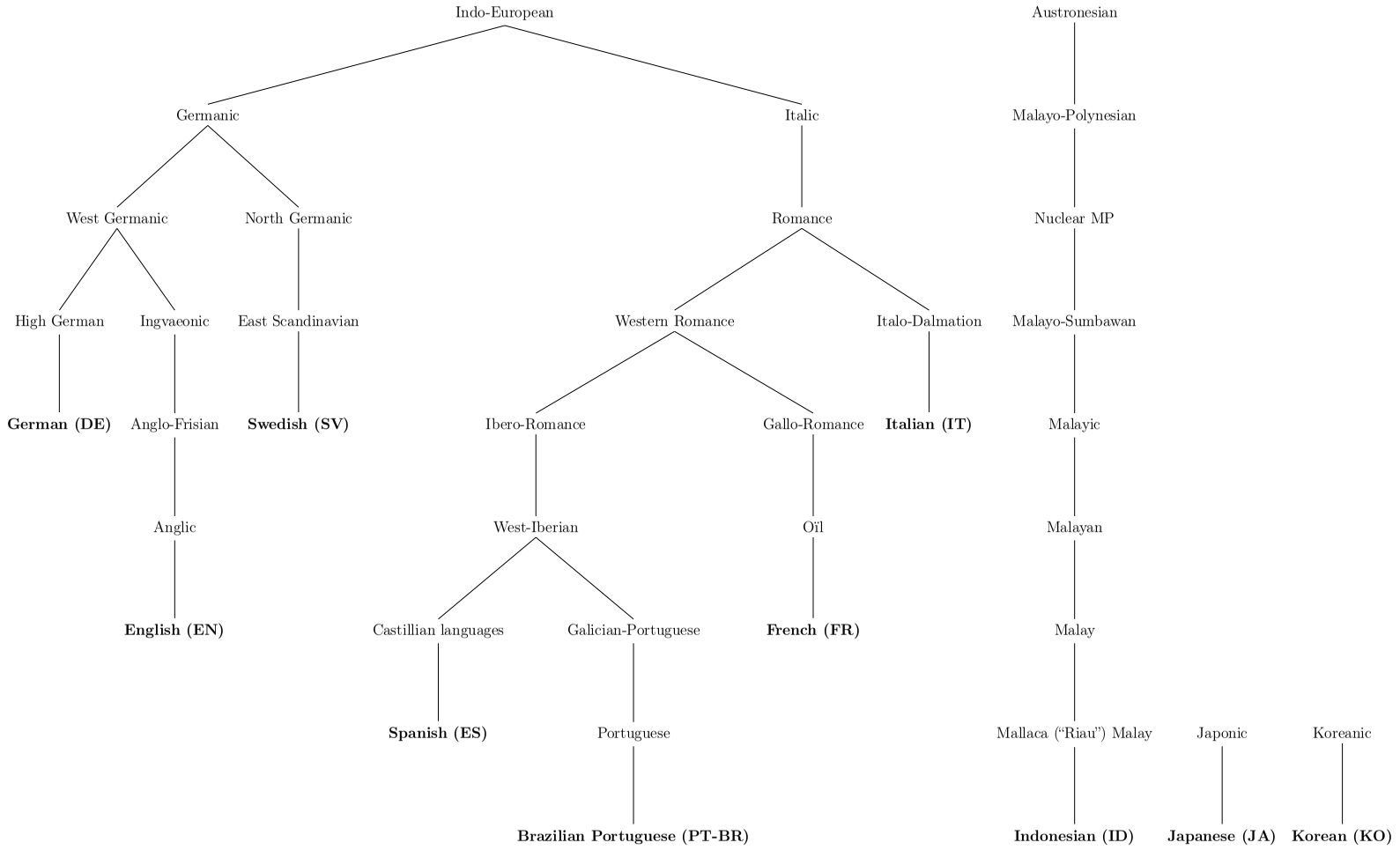

Figure 3.1: A diagram showing how the ten languages used are related.

maximum spanning trees over directed graphs, using models based on large-margin discriminative training methods. I use default parameters across the board.

This package includes automatic calculation of evaluation metrics, including unlabeled attachment scores (UAS) and labeled attachment scores (LAS). The unlabeled attachment score of a dependency parsed sentence is equal to the number of words that depend on the correct head, divided by the total number of words. By contrast, the labeled attachment score also requires that the dependency have the correct label, that is, it is equal to the number of words that depend on the correct head with the correct label, divided by the total number of words. Note that punctuation is often excluded in reported results, but is included in the MSTParser's calculations. All results reported use gold POS tags for the test data; incorporating automatic POS tagging into the pipeline for testing is left to future work.

3.2.1 Baseline and Upperbound

For my baseline and upperbound, I use the direct transfer technique. For a given language, its upperbound is the accuracy when tested on a system trained on the same language, and its baselines are the accuracies when tested on systems trained on each of the other languages. I also present multi-source language baselines, where for each language, I train on the remaining nine languages in the Google Universal Dependency Treebank. Additionally, in order to directly compare my instance selection results, I present random selection baselines, where a set number of sentences are selected randomly from the pool of possible training data. Since the MSTParser uses an online learning algorithm with multiple iterations, the order of the training data has an impact on the resulting score. For this reason, all baseline and upperbound experiments use the average of five trials with varying random order of the training data (and varying randomly selected training instances, in the case of the random selection baselines) to account for any variance as a result of order or selection.

3.2.2 Instance Selection

I first attempt to recreate the methodology from Søgaard (2011) for instance selection. For each language, I use the POS tag sequences from their training data to train a backoff trigram language model using SRILM, freely available online.[4] No unknown unigram handling is needed, as all 17 universal POS tags occur in every training data set. For each of these, I then use the perplexity per word for the POS sequences for each sentence in the training data from the other nine languages to rank the sentences by their similarity to the target language. The most similar sentences are used to train a dependency parser. My methodology differs slightly from that of Søgaard (2011) as I opt to select a set number of sentences, rather than a percentage of the training data available. I also experiment with using a set threshold for the perplexity, which results in training sets that vary drastically in size, depending on the test language.

I then apply the method of cross-entropy difference described in Moore and Lewis (2010) and implemented for machine translation domain adaptation in Axelrod et al. (2011). For this, I use the language model trained on the POS tag sequences from the target language and additionally train another backoff trigram language model on the POS tag sequences of the pool of training data from the remaining nine source languages. From the POS sequence for each possible training sentence, the cross-entropy is calculated with regards to both language models and the difference is taken to rank the sentences. Once again, I select a set number of sentences and try a set threshold in order to determine the optimal pool of training data.

Using multiple source languages allows for the possibility of selecting training data that is closer to the test data, since relevant training data for Brazilian Portuguese is more likely to be found in Spanish, while relevant training data for Japanese is more likely to be found in Korean. However, a pool of training data in multiple languages may not always be available. Additionally, training a language model on multiple

[4]http://www.speech.sri.com/projects/srilm/

languages, even abstracted to POS tags, raises some potential concerns; each language is likely to have different sequences and different probabilities for these sequences, and pooling all the languages together means that these probabilities are conflated. For these reasons, I also repeat the previous experiments using only English as the source language. As previously mentioned, English has by far the largest quantity of training data. This is clearly not a quirk of this particular dataset; in general, training data in English is much more prevalent and readily available for a variety of tasks, so using English allows for a methodology that may be easily adapted and tested for other scenarios.

Since, as mentioned, relevant training data for a particular language may not be as likely to be found in English, I also explore the possibility of automatically rearranging POS tags in the English training data in order to better fit the test language. Because of the scope of this work, this is confined to rearrangement of adjectives and nouns. English typically has adjectives preceding nouns (note, though, that in the training data set used, there are already a few, although infrequent, occurrences of an adjective following a noun but still modifying it). In order to improve the performance of a parser trained only on English for languages where adjectives are much more likely to follow the noun that they modify, I rearrange all occurrences of adjectives preceding the noun they modify so that the adjectives occur after the noun, while maintaining all other ordering and dependencies. I then test on the three languages which have the most occurrences of modifying adjectives following the noun relative to modifying adjectives preceding the noun; these are Indonesian, Spanish, and French, where over 70% of modifying adjectives occur following the noun rather than preceding it (in the case of Indonesian, this is actually over 90%). Using this modified English training set, I then repeat the instance selection experiments.

language	# of sentences in training set	# of tokens in training set	average sentence length	# of sentences in test set	# of tokens in test set	average sentence length
DE	14,118	264,906	18.76	1,000	16,339	16.34
EN	39,832	950,028	23.85	2,416	56,684	23.46
SV	4,447	66,631	14.98	1,219	20,376	16.72
ES	14,138	375,180	26.54	300	8,295	27.65
FR	14,511	351,233	24.20	300	6,950	23.17
IT	6,389	149,145	23.34	400	9,187	22.97
PT-BR	9,600	239,012	24.90	1,198	29,438	24.57
ID	4,477	97,531	21.78	557	11,780	21.15
JA	8,277	80,172	9.69	299	2,760	9.23
KO	5,437	60,621	11.15	299	2,631	8.80

Table 3.1: Data Statistics for Google Universal Dependency Treebank v2.0

Chapter 4

RESULTS

The results for all experiments are detailed below. Unlabeled and labeled attachment scores are reported. First, the baseline and upperbound numbers are presented. Then, results for instance selection methods on a training data pool of multiple source languages are discussed. Next, results for the same methods on a training data pool of English training data only are shown. Lastly, results for rearranging the English training data and using instance selection on the rearranged data are presented.

4.1 Baseline and Upperbound

The single-source baseline and upperbound results are shown below in Tables 4.1 and 4.2. In italics are the upperbound scores, where the source and target language are the same. In bold is the best non-same source language for each target language. These results are similar to the direct transfer results shown by McDonald et al. (2013), although a different dependency parser is utilized (in their case, a perceptron-trained shift-reduce parser, while in my case, the MSTParser).

Notably, the upperbound for UAS is in the upper 70s for all the languages, with the exception of Korean which is slightly lower (perhaps not surprising given that Korean is the language with the smallest training data set), but still in the 70s, and English which exceeds 80%, likely because of its very large training data set. The best non-same source language is typically a highly related language - when testing on a Germanic language, training on one of the other Germanic languages tends to give the best results and the same holds for the Romance languages. In some cases, these scores come rather close to the upperbound, specifically for the Romance languages

where the best non-same language (always another Romance language) yields a score within about four percentage points of training on the language itself. For Japanese, the best non-same language is Korean, and vice versa, and the UAS is quite high (similar to the best non-same language UAS for the Germanic languages), especially given that these are not related languages. Indonesian is a noticeable outlier with the best non-same UAS (the result of training on Spanish) at only 50.37%.

For all languages, the LAS is predictably lower, although by how much varies greatly. The upperbound for LAS is 65% and 75% for all languages, with similar exceptions to before - an exceptionally high score for English and an exceptionally low score for Korean (below 60%), probably due to the variance in training data quantity. The best non-same source language is the same for the LAS as for the UAS for every language except French, where Spanish as a source language goes from having a slightly lower UAS score than Italian to having a slightly higher LAS. Eight of the ten languages have a best non-same source language LAS score that is around 10 percentage points lower than the best non-same source language UAS score. For Japanese and Korean, however, this score is around 40 percentage points lower. So Japanese parsed with a dependency parser trained on Korean has a UAS of 71.54%, but a LAS of 30.35%. This drop indicates that while the parser is correctly assigning which word the words depend on, over half the time, it is doing so with an incorrect label. This may be because of inconsistent annotation or because of similar POS tag sequences and dependency directionality having different grammaticality in the two languages.

The multi-source baseline results are shown in Tables 4.3 and 4.4, where for each target language, all the available training data, in one condition, and the training data from the remaining nine languages, in another, is combined to be used as the source.

For all languages, a drop in both UAS and LAS occurs when all the training data is pooled together, even when the training data from the target language is included

in that pool. The amount of this drop varies by language; unsurprisingly, English only sees a very small drop, likely because of the vast quantity of English training data, while most of the other languages see a drop of about 5%. The drop for Indonesian and Japanese is larger, with Japanese experiencing nearly a 10% drop in UAS and over 20% drop in LAS, while Indonesian has over 25% drop for both UAS and LAS. This provides justification to the idea that having more data isn't always better; especially in the case of these languages, having extra training data that is irrelevant data (i.e. from other languages) has a significant negative impact on the dependency parses for these languages.

In fact, for many of the languages, including all the Romance languages, Swedish, and Indonesian, removing the training data for the language itself only drops the score a few more percentage points, indicating that much of the impact for the dependency parses is not coming from the highly relevant in-language training data even when it's included.

Source Training Language	Target Test Language									
	Unlabeled Attachment Score (UAS)									
	Germanic			Romance						
	DE	**EN**	**SV**	**ES**	**FR**	**IT**	**PT-BR**	**ID**	**JA**	**KO**
DE	*76.04%*	50.59%	64.38%	57.22%	57.94%	57.09%	58.97%	39.47%	38.43%	39.91%
EN	57.15%	*83.90%*	**70.00%**	65.30%	68.44%	68.00%	65.82%	42.40%	41.14%	40.97%
SV	**60.25%**	57.52%	*79.41%*	64.51%	63.95%	63.56%	61.73%	45.12%	38.83%	35.01%
ES	55.93%	58.91%	67.97%	*76.48%*	72.74%	74.25%	**75.56%**	**50.37%**	35.70%	35.84%
FR	56.25%	**61.07%**	68.74%	71.37%	*77.27%*	**74.63%**	72.00%	46.72%	32.11%	33.98%
IT	55.11%	59.95%	68.94%	71.63%	**72.86%**	*79.43%*	71.82%	48.48%	35.00%	35.48%
PT-BR	57.23%	58.89%	68.02%	**73.78%**	72.25%	74.01%	*79.03%*	49.78%	34.52%	35.23%
ID	54.31%	42.09%	59.03%	62.99%	63.51%	62.53%	63.27%	*77.76%*	23.53%	23.53%
JA	29.36%	29.61%	25.77%	25.83%	26.16%	23.02%	21.82%	16.46%	*79.88%*	**67.88%**
KO	32.10%	33.60%	28.49%	27.92%	27.53%	24.21%	24.01%	16.74%	**71.54%**	*73.04%*

Table 4.1: Baseline and upperbound unlabeled attachment scores (UAS)

Source Training Language	Target Test Language									
	Labeled Attachment Score (LAS)									
	Germanic			Romance						
	DE	**EN**	**SV**	**ES**	**FR**	**IT**	**PT-BR**	**ID**	**JA**	**KO**
DE	*67.06%*	42.92%	52.26%	47.89%	48.20%	48.20%	52.18%	29.79%	13.09%	26.10%
EN	48.54%	*79.38%*	**59.26%**	55.16%	58.27%	58.55%	59.65%	36.33%	9.77%	24.72%
SV	**52.37%**	48.04%	*71.08%*	53.35%	53.26%	53.78%	53.82%	33.60%	12.36%	19.66%
ES	47.43%	50.04%	54.76%	*68.45%*	**63.30%**	64.77%	**69.99%**	**41.51%**	11.79%	22.47%
FR	48.17%	**52.40%**	55.62%	61.77%	*69.54%*	**64.98%**	66.12%	37.14%	9.33%	20.78%
IT	45.03%	45.63%	54.29%	60.79%	61.26%	*72.58%*	62.43%	38.89%	11.15%	14.44%
PT-BR	48.68%	50.75%	54.77%	**65.30%**	62.74%	63.86%	*74.02%*	40.67%	11.04%	22.11%
ID	44.19%	35.37%	47.54%	52.69%	52.96%	54.29%	56.00%	*70.86%*	12.30%	19.22%
JA	7.51%	3.96%	5.04%	2.88%	3.61%	3.76%	3.05%	6.68%	*66.78%*	**30.43%**
KO	25.72%	24.01%	18.55%	19.19%	18.66%	15.74%	17.05%	12.66%	**30.35%**	*58.18%*

Table 4.2: Baseline and upperbound labeled attachment scores (LAS)

Source Training Language	Target Test Language									
	Unlabeled Attachment Score (UAS)									
	Germanic			Romance						
	DE	EN	SV	ES	FR	IT	PT-BR	ID	JA	KO
test language training data	76.04%	83.90%	79.41%	76.48%	77.27%	79.43%	79.03%	77.76%	79.88%	73.04%
all (including test language)	68.58%	81.40%	74.85%	73.74%	75.96%	76.04%	74.63%	51.19%	70.36%	67.50%
all (excluding test language)	58.88%	59.14%	73.15%	71.90%	73.81%	74.54%	72.53%	45.64%	57.55%	60.43%

Table 4.3: Baseline and upperbound unlabeled attachment scores (UAS) with multiple source languages

Source Training Language	Target Test Language									
	Labeled Attachment Score (LAS)									
	Germanic			Romance						
	DE	EN	SV	ES	FR	IT	PT-BR	ID	JA	KO
test language training data	67.06%	79.38%	71.08%	68.45%	69.54%	72.58%	74.02%	70.86%	66.78%	58.18%
all (including test language)	60.16%	76.23%	64.62%	64.82%	66.79%	66.71%	69.12%	44.38%	44.04%	47.26%
all (excluding test language)	51.19%	51.73%	62.78%	62.27%	64.21%	65.18%	66.82%	39.08%	22.10%	34.78%

Table 4.4: Baseline and upperbound labeled attachment scores (LAS) with multiple source languages

4.2 Instance Selection

In the following experiments, methods for selecting less, but hopefully better, training data from the pool of available training data are tested.

4.2.1 Multiple Source Languages

For instance selection using multiple source languages, note that the training data for each target language is considered to be the pool of training data from every language except for that language. Since each language has a slightly different pool of training data, the quantity of the full set of training data differs, with English having the least training data (since its own very large training data set is omitted) with around 81,000 sentences and Swedish having the most training data with a little less than 117,000 sentences.

Random Selection Baseline

The random selection baseline results for selecting training data from multiple source languages are shown in Tables 4.5 and 4.6. As one would expect, for all languages, the performance improves as more sentences are used for training. However, the extent of the range between the lowest scores, when only using 1,000 randomly selected sentences for training, and the highest scores, when using the full set of combined training data differs by language. For German, this range is less than 2%, for both UAS and LAS, while for Indonesian, the range is a little over 2%, and for English, the range is around 3%. For the Romance languages and Swedish, the range is around 4%. Interestingly, Japanese and Korean have a much bigger range for UAS than for LAS with a range of close to 6% for UAS for Japanese compared to about 4% in LAS, while the range for Korean is over 8% for UAS but under 5% for LAS.

Perplexity Selection

Results for selecting a set number of sentences as training data based on their perplexity against a trigram language model trained on the POS tags of the test language's training data are shown in Tables 4.7 and 4.8. While using all of the training data still gives the best results for some languages (notably, all the Germanic languages), using less training data selected this way does provide a boost in scores for other languages.

For the Germanic languages, the results achieved when using perplexity to select a set number of sentences are worse than random selection for a smaller number of sentences and comparable when selecting more sentences. In the case of German, equivalence to random selection happens around 10,000 sentences, while for English, this is around 20,000 sentences, and for Swedish, although the results get closer to random selection numbers, they stay slightly worse even when selecting 20,000 sentences. For these languages, there is no point at which selecting more sentences using perplexity gives results better than simply using all the training data.

Among the Romance languages, French and Italian follow a similar trend to the Germanic languages, with worse results than random selection for a smaller number of sentences (1,000 or 5,000) and comparable results when selecting more. These languages do achieve a slightly higher UAS when selecting 20,000 sentences using perplexity compared to selecting 20,000 sentences randomly, but it is less than 1% higher. Although Spanish and Brazilian Portuguese also have worse results than random selection for a smaller number of sentences, their scores, both UAS and LAS, exceed random selection by between 1.5% and 2.5% when selecting 10,000 or 20,000 sentences. Compared to using all the data, there is no advantage for French or Italian, but Spanish and Brazilian Portuguese achieve slightly higher scores, with a .53% boost in UAS and 1.06% boost in LAS for Spanish and 1.49% boost in UAS and 1.65% boost in LAS for Brazilian Portuguese, when 20,000 sentences are selected using perplexity.

Indonesian has worse scores than random selection when selecting 1,000 sentences using perplexity, but scores when selecting 5,000 or 10,000 sentences are slightly higher (by around .6%) for UAS and comparable for LAS. There is a boost of close to 1.5% in UAS and LAS when selecting 20,000 sentences using perplexity compared to randomly selecting 20,000 sentences. However, those scores are still not significantly better than using all the training data.

Japanese and Korean follow a dramatically different trend than the other languages. Both achieve very large increases in UAS over random selection when using perplexity selection for 1,000, 5,000, or 10,000 sentences, although selecting 20,000 sentences using perplexity provides results slightly worse than randomly selecting 20,000 sentences. The increase over random is highest at 1,000 sentences where Japanese achieves a UAS 15.38% higher than random selection and Korean achieves a UAS 12% higher than random selection. For Japanese, a similar trend is seen with the LAS, although the boost over random selection is not quite as much, with 1,000 sentences selected by perplexity giving a LAS 10.77% higher than randomly selecting 1,000 sentences. For Korean, however, the LAS when using perplexity to select the sentences is lower than random selection, by as much as 4.6% (when selecting 10,000 sentences). For both languages, perplexity selection give a large boost in UAS over using all the training data, with Japanese achieving a UAS 9.55% higher than using all the data, when 1,000 sentences are selected, and Korean achieving a UAS 6.74% higher than using all the data, when 5,000 sentences are selected. Although Korean doesn't show a boost in LAS over using all the data, Japanese achieves a LAS 6.74% higher than using all the data, when 1,000 sentences are selected.

These promising results raise the question of whether or not consistent results can be achieved across this variety of languages. How can one possibly know, a priori, whether selecting 1,000 sentences (as in the case of Japanese) or 20,000 sentences (as in the case of Brazilian Portuguese) or using all of the data available (as in the case of the Germanic languages) will provide the best results? While fairly good results can

be achieved across the languages by selecting 5,000 sentences, this method still results in scores that vary from the best possible scores by as much as 3%. This creates the motivation for determining a potential perplexity cut-off - a threshold where using those sentences in the training data below the threshold results in a higher score than the score achieved incorporating sentences above the threshold. Naturally, the number of training data sentences below a given threshold will vary drastically by language and be impacted significantly by the similarity of the target language to the source languages.

Results for selecting sentences as training data based on those that are below a certain cut-off for perplexity against a trigram language model trained on the POS tags of the test language's training data are shown in Tables 4.9 and 4.10. They are also presented in Figures 4.1 and 4.2. While the cut-off that results in the highest score varies from language to language, there seems to be a very promising trend. All languages achieve a UAS within 1% of their highest UAS when a cut-off of 6 is used. This is similarly true for the LAS, with the sole exception of Korean whose LAS is still more than 4% lower than the best, which is achieved when using all the training data. That is to say, given a new target language and a trigram language model trained on POS tag sequences in that language, one can reliably select all training data sentences from the available pool of training data that are below a perplexity of 6 relative to that language model to achieve a UAS and LAS that are very close to the best possible. In the case of Japanese, this is an over 10% boost in UAS over using all of the training data.

Number of Sentences	Target Test Language									
	Unlabeled Attachment Score (UAS)									
	Germanic			Romance						
	DE	EN	SV	ES	FR	IT	PT-BR	ID	JA	KO
1,000	57.32%	56.90%	69.43%	68.37%	69.41%	70.31%	69.13%	43.48%	51.72%	52.35%
5,000	58.27%	58.06%	70.82%	69.60%	71.32%	72.15%	70.60%	44.39%	53.73%	56.35%
10,000	58.33%	58.36%	71.56%	70.25%	72.39%	73.13%	71.12%	44.82%	54.84%	58.27%
20,000	58.33%	58.79%	72.18%	70.86%	72.81%	73.68%	71.68%	44.60%	55.80%	59.41%
all	58.88%	59.14%	73.15%	71.90%	73.81%	74.54%	72.53%	45.64%	57.55%	60.43%

Table 4.5: Unlabeled attachment scores (UAS) for random selection from multiple source languages (average of five trials)

Number of Sentences	Target Test Language									
	Labeled Attachment Score (LAS)									
	Germanic			Romance						
	DE	EN	SV	ES	FR	IT	PT-BR	ID	JA	KO
1,000	49.53%	48.70%	58.92%	58.83%	59.93%	61.03%	63.36%	36.93%	18.07%	30.16%
5,000	50.39%	50.56%	60.62%	60.19%	61.98%	62.94%	64.95%	37.73%	19.27%	31.78%
10,000	50.68%	51.08%	61.22%	60.76%	63.15%	63.91%	65.42%	38.21%	20.30%	33.87%
20,000	50.69%	51.38%	61.82%	61.26%	63.34%	64.51%	66.04%	38.05%	21.10%	34.17%
all	51.19%	51.73%	62.78%	62.27%	64.21%	65.18%	66.82%	39.08%	22.10%	34.78%

Table 4.6: Labeled attachment scores (LAS) for random selection from multiple source languages (average of five trials)

Number of Sentences	Target Test Language									
	Unlabeled Attachment Score (UAS)									
	Germanic			Romance						
	DE	**EN**	**SV**	**ES**	**FR**	**IT**	**PT-BR**	**ID**	**JA**	**KO**
1,000	55.32%	54.44%	64.27%	63.89%	67.57%	68.59%	65.52%	40.38%	**67.10%**	64.35%
5,000	57.77%	56.97%	70.01%	70.07%	70.91%	71.84%	70.73%	45.07%	66.63%	**66.90%**
10,000	58.39%	57.09%	70.56%	71.86%	72.65%	73.13%	72.92%	45.43%	60.87%	64.69%
20,000	58.64%	58.44%	71.44%	**72.43%**	73.34%	**74.61%**	**74.02%**	**46.07%**	55.25%	58.65%
all	**58.88%**	**59.14%**	**73.15%**	71.90%	**73.81%**	74.54%	72.53%	45.64%	57.55%	60.43%

Table 4.7: Unlabeled attachment scores (UAS) for perplexity selection from multiple source languages

Number of Sentences	Target Test Language									
	Labeled Attachment Score (LAS)									
	Germanic			Romance						
	DE	**EN**	**SV**	**ES**	**FR**	**IT**	**PT-BR**	**ID**	**JA**	**KO**
1,000	46.52%	45.63%	52.01%	54.58%	57.30%	58.62%	59.46%	33.46%	**28.84%**	27.82%
5,000	49.73%	50.03%	58.86%	60.66%	61.09%	61.86%	64.73%	37.36%	27.17%	30.10%
10,000	50.43%	50.47%	59.50%	62.47%	62.82%	63.44%	67.27%	38.35%	23.77%	29.27%
20,000	50.74%	51.37%	60.72%	**63.33%**	63.47%	64.78%	**68.47%**	**39.30%**	20.69%	34.59%
all	**51.19%**	**51.73%**	**62.78%**	62.27%	**64.21%**	**65.18%**	66.82%	39.08%	22.10%	**34.78%**

Table 4.8: Labeled attachment scores (LAS) for perplexity selection from multiple source languages

Perplexity Cut-Off	Target Test Language									
	Unlabeled Attachment Score (UAS)									
	Germanic			Romance						
	DE	**EN**	**SV**	**ES**	**FR**	**IT**	**PT-BR**	**ID**	**JA**	**KO**
3	43.77%	37.85%	49.39%	59.55%	62.86%	61.34%	61.92%	26.40%	62.72%	62.03%
4	57.06%	56.84%	67.39%	70.27%	71.12%	72.05%	71.15%	39.18%	67.61%	65.83%
5	58.75%	57.41%	70.77%	72.26%	73.51%	74.54%	**74.21%**	45.31%	67.79%	**66.63%**
6	58.75%	58.82%	72.23%	**72.73%**	74.33%	**75.25%**	73.94%	45.86%	**68.88%**	66.36%
7	59.05%	**59.85%**	72.60%	72.44%	74.13%	74.69%	73.65%	45.87%	67.46%	65.98%
8	**59.11%**	59.47%	72.99%	72.68%	74.23%	74.61%	72.71%	**46.27%**	67.54%	65.26%
9	59.09%	59.80%	73.02%	72.41%	**74.42%**	74.44%	72.70%	46.18%	66.63%	65.49%
none	58.88%	59.14%	**73.15%**	71.90%	73.81%	74.54%	72.53%	45.64%	57.55%	60.43%

Table 4.9: Unlabeled attachment scores (UAS) for perplexity selection from multiple source languages

Perplexity Cut-Off	Target Test Language									
	Labeled Attachment Score (LAS)									
	Germanic			Romance						
	DE	**EN**	**SV**	**ES**	**FR**	**IT**	**PT-BR**	**ID**	**JA**	**KO**
3	29.94%	20.37%	35.95%	46.53%	51.54%	51.28%	54.61%	16.72%	25.94%	26.99%
4	48.91%	48.82%	55.64%	61.22%	61.09%	62.38%	65.34%	32.39%	**29.24%**	29.46%
5	50.82%	50.62%	59.56%	63.33%	63.68%	64.93%	**68.62%**	38.31%	28.73%	29.91%
6	50.97%	51.85%	61.69%	**63.82%**	64.59%	**65.70%**	68.32%	39.05%	29.06%	30.03%
7	51.23%	52.74%	62.07%	63.24%	64.42%	65.33%	68.08%	39.18%	28.12%	29.27%
8	51.34%	52.48%	62.57%	63.34%	64.65%	65.29%	67.08%	39.58%	27.68%	29.00%
9	**51.52%**	**52.81%**	62.53%	62.85%	**64.71%**	65.22%	67.05%	**39.74%**	27.43%	29.38%
none	51.19%	51.73%	**62.78%**	62.27%	64.21%	65.18%	66.82%	39.08%	22.10%	**34.78%**

Table 4.10: Labeled attachment scores (LAS) for perplexity selection from multiple source languages

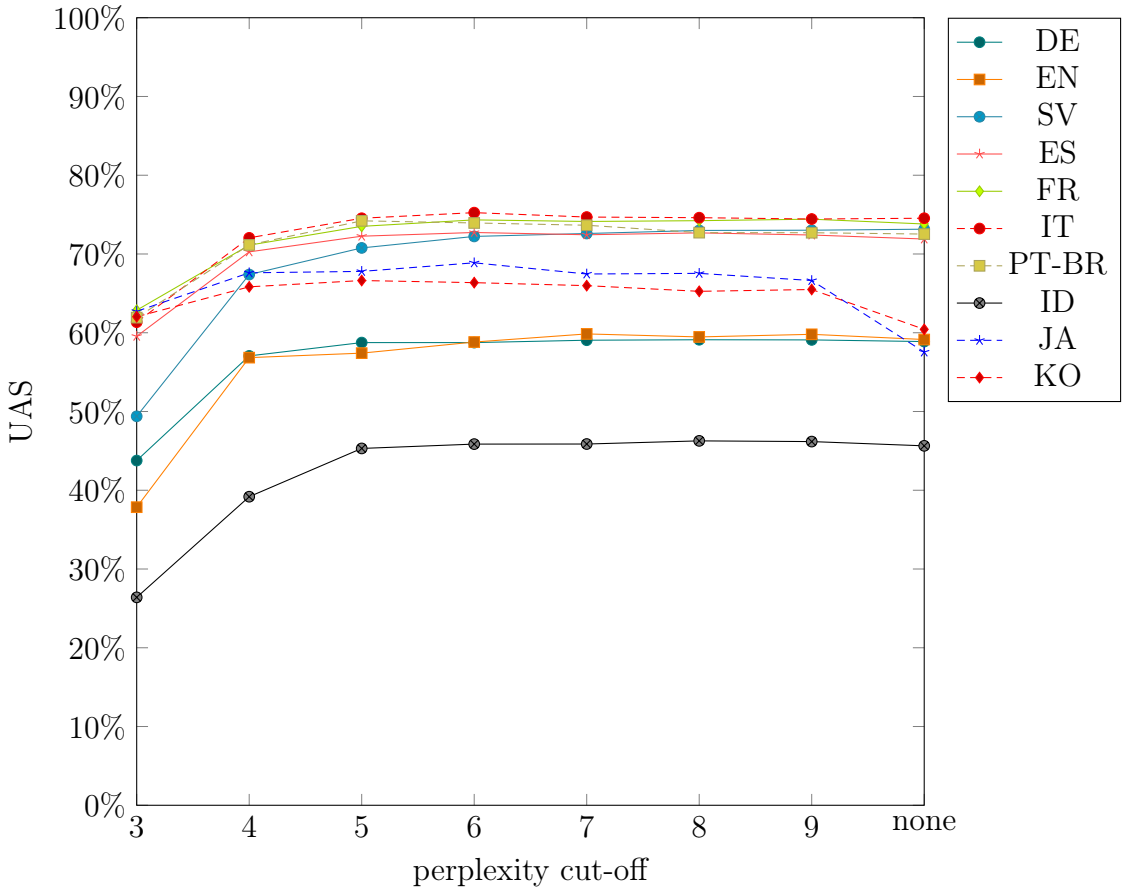

Figure 4.1: Unlabeled attachment scores (UAS) with different perplexity thresholds for selecting training data from multiple source languages

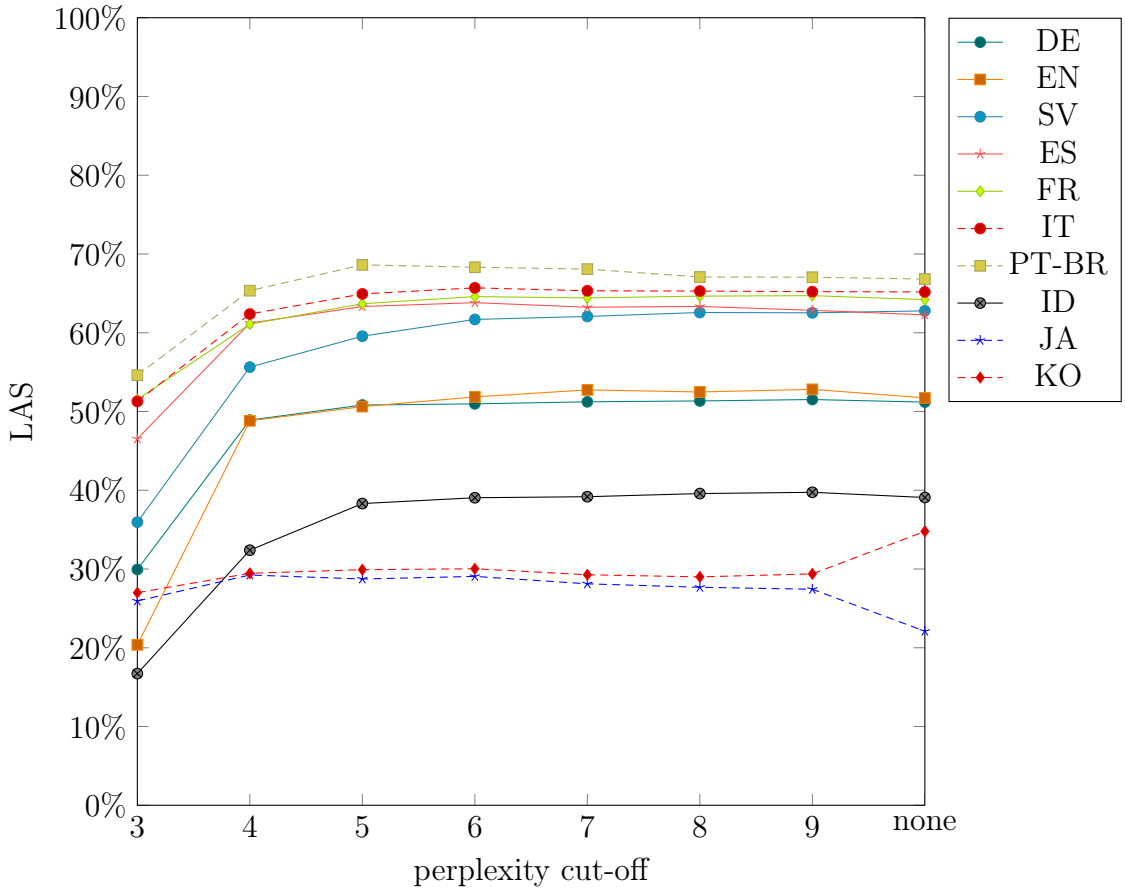

Figure 4.2: Labeled attachment scores (LAS) with different perplexity thresholds for selecting training data from multiple source languages

Cross-Entropy Difference Selection

While perplexity selection has promising results, perhaps another technique from MT domain adaptation can be leveraged to achieve even higher scores. Cross-entropy difference selection introduces the use of another language model in addition to the one trained on the target language. The results for selecting a set number of sentences from the training data based on cross-entropy difference are shown in Tables 4.11 and 4.12. Although the trends are similar to perplexity selection, most of the languages achieve better results with cross-entropy difference selection than with perplexity selection.

For the Germanic languages, the results closely resemble those of selecting using perplexity. However, German actually achieves slightly (less than 1%) better scores than random selection for each number of sentences. English and Swedish still have scores worse than random when selecting 1,000 sentences, but comparable LAS to random selection are achieved when selecting 5,000 or more sentences, with comparable UAS at these number of sentences for Swedish as well, although English UAS are worse than random selection for all quantities of sentences. Once again, no significant boost is seen for any of the languages over using all the training data, although German does achieve very slightly (less than .5%) higher scores when 20,000 sentences are selected using cross-entropy difference.

The Romance languages all achieve better results than with perplexity selection. The scores, both UAS and LAS, when selecting 1,000 sentences are equivalent or slightly (less than 1%) better than random selection, and the scores when selecting more sentences are all better than random selection by between approximately 1% and 3.5%. French and Italian achieve a boost of a little more than .5% over using all the training data when selecting 20,000 (French) or 10,000 (Italian) sentences. Spanish and Brazilian Portuguese see an even bigger boost, with Spanish achieving a 1.6% increase in UAS and 2.17% increase in LAS over using all the training data

when selecting 20,000 sentences and Brazilian Portuguese achieving a 2.12% increase in UAS and 2.05% increase in LAS over using all the training data when selecting 20,000 sentences.

Although the highest UAS and LAS achieved by cross-entropy difference selection for Indonesian are not higher than the highest scores achieved by perplexity selection, the scores when selecting 1,000 sentences are higher. Otherwise, the results appear very similar to perplexity selection, and there is once again no boost over using all the training data.

Japanese and Korean follow the same trend as with perplexity selection, with very large improvements in UAS over random selection when selecting 1,000, 5,000, or 10,000 sentences. Similarly, selecting 20,000 sentences using cross-entropy difference provides results worse than randomly selecting 20,000 sentences. The increase over random is once again highest at 1,000 sentences, where Japanese achieves a UAS 16.43% higher than random selection and Korean achieves a UAS 13.79% higher than random selection; note that these scores are over 1% better than those with perplexity selection. Again, the same trend is seen in LAS for Japanese, with 1,000 sentences selected by cross-entropy difference giving a LAS 10.95% higher than randomly selecting 1,000 sentences, although this is not significantly better than the boost seen with perplexity selection, while for Korean, the LAS when using cross-entropy difference to select the sentences is lower than random selection, by as much as 4.41%. However, Korean does achieve a higher LAS than random selection when selecting 20,000 sentences, with a boost of 2.13% over random selection. For both languages, cross-entropy difference selection gives a boost over using all the training data, for both UAS and LAS. For Japanese, this is a 10.60% boost in UAS (more than 1% better than perplexity selection) and 6.92% boost in LAS (comparable to perplexity selection). For Korean, this is a 5.93% boost in UAS (less than 1% worse than perplexity selection) and a 1.52% boost in LAS (compared to no boost with perplexity selection).

These results seem to imply that cross-entropy difference selection is overall a better method of instance selection than perplexity selection, although in most cases, the difference is not a very large one.

The results for selecting from the training data based on a cross-entropy difference threshold are shown in Tables 4.13 and 4.14 and Figures 4.3 and 4.4. Similar to the results with using a perplexity threshold, these results show a promising trend; by using a cross-entropy difference cut-off of 0 to select training data, most languages achieve a UAS within just .30% of their highest UAS. Exceptions are English, Swedish, and Indonesian, where the difference is a little over 1%. There is a little more variance in how close the LAS with this cut-off is to the highest achieved; for most languages, it is within 1% of the highest LAS. However, Japanese has a difference a little over 1% between the LAS with a cross-entropy difference cut-off of 0 and the highest LAS. Most notably, Korean shows the biggest deviation with the highest LAS being over 5% higher than that achieved with a cross-entropy difference cut-off of 0, despite having the highest UAS at that cut-off.

Number of Sentences	Target Test Language									
	Unlabeled Attachment Score (UAS)									
	Germanic			Romance						
	DE	EN	SV	ES	FR	IT	PT-BR	ID	JA	KO
1,000	58.04%	54.50%	65.87%	68.96%	70.40%	69.87%	69.38%	44.20%	**68.15%**	66.14%
5,000	58.84%	56.89%	70.38%	71.67%	72.95%	74.36%	73.20%	44.98%	67.86%	**66.36%**
10,000	59.04%	57.36%	71.48%	73.27%	73.18%	**75.10%**	73.80%	45.02%	61.09%	65.91%
20,000	**59.30%**	58.00%	72.71%	**73.50%**	**74.37%**	75.03%	**74.65%**	45.63%	54.42%	57.62%
all	58.88%	**59.14%**	**73.15%**	71.90%	73.81%	74.54%	72.53%	**45.64%**	57.55%	60.43%

Table 4.11: Unlabeled attachment scores (UAS) for cross-entropy difference selection from multiple source languages

Number of Sentences	Target Test Language									
	Labeled Attachment Score (LAS)									
	Germanic			Romance						
	DE	**EN**	**SV**	**ES**	**FR**	**IT**	**PT-BR**	**ID**	**JA**	**KO**
1,000	50.16%	47.42%	55.50%	59.30%	60.81%	60.73%	63.18%	36.06%	**29.02%**	29.19%
5,000	50.77%	50.89%	60.16%	62.83%	63.25%	65.04%	67.31%	37.88%	27.61%	29.42%
10,000	51.32%	51.02%	61.44%	64.28%	63.90%	65.90%	68.07%	38.53%	24.09%	29.46%
20,000	**51.51%**	51.59%	62.39%	**64.44%**	**64.81%**	**65.55%**	**68.87%**	38.93%	20.73%	**36.30%**
all	51.19%	**51.73%**	**62.78%**	62.27%	64.21%	65.18%	66.82%	**39.08%**	22.10%	34.78%

Table 4.12: Labeled attachment scores (LAS) for cross-entropy difference selection from multiple source languages

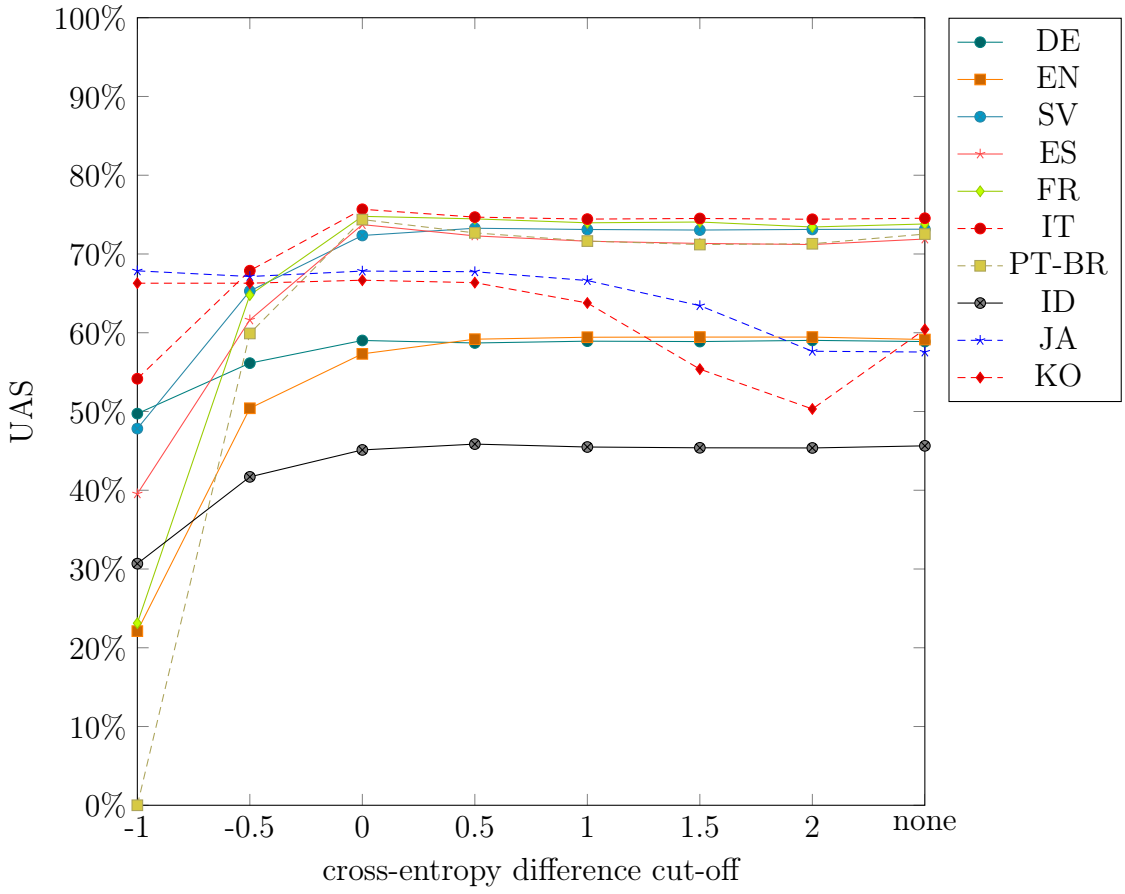

Figure 4.3: Unlabeled attachment scores (UAS) with different cross-entropy difference thresholds for selecting training data from multiple source languages

Cross-Entropy Difference Cut-Off	Target Test Language									
	Unlabeled Attachment Score (UAS)									
	Germanic			Romance						
	DE	EN	SV	ES	FR	IT	PT-BR	ID	JA	KO
-1	49.73%	22.11%	47.83%	39.55%	23.12%	54.16%	n/a*	30.68%	**67.86%**	66.29%
-0.5	56.14%	50.41%	65.30%	61.63%	64.78%	67.88%	59.91%	41.70%	67.14%	66.29%
0	59.02%	57.32%	72.35%	**73.73%**	**74.78%**	**75.68%**	**74.37%**	45.12%	67.83%	**66.67%**
0.5	58.70%	59.18%	73.27%	72.30%	74.46%	74.68%	72.68%	**45.86%**	67.75%	66.36%
1	58.93%	59.43%	73.11%	71.62%	73.96%	74.43%	71.64%	45.49%	66.63%	63.78%
1.5	58.88%	**59.45%**	73.04%	71.34%	74.06%	74.51%	71.20%	45.39%	63.44%	55.38%
2	**59.03%**	**59.45%**	73.12%	71.20%	73.43%	74.41%	71.30%	45.37%	57.65%	50.32%
none	58.88%	59.14%	**73.15%**	71.90%	73.81%	74.54%	72.53%	45.64%	57.55%	60.43%

*No training data met this cut-off.

Table 4.13: Unlabeled attachment scores (UAS) for cross-entropy difference selection from multiple source languages

Cross-Entropy Difference Cut-Off	Target Test Language									
	Labeled Attachment Score (LAS)									
	Germanic			Romance						
	DE	EN	SV	ES	FR	IT	PT-BR	ID	JA	KO
-1	38.63%	12.21%	31.58%	23.99%	11.24%	44.49%	n/a*	18.68%	**29.24%**	28.85%
-0.5	47.93%	43.12%	54.91%	52.26%	54.65%	58.82%	53.01%	33.93%	28.37%	29.84%
0	51.25%	51.14%	62.21%	**64.70%**	**65.01%**	**66.29%**	**68.75%**	38.24%	28.01%	29.50%
0.5	51.23%	51.56%	**62.81%**	62.56%	64.84%	65.07%	66.94%	39.25%	27.25%	29.50%
1	51.26%	51.66%	62.63%	61.80%	64.37%	64.92%	65.89%	39.01%	27.07%	31.24%
1.5	51.23%	51.49%	62.62%	61.44%	64.42%	64.93%	65.46%	38.79%	25.62%	34.59%
2	**51.36%**	51.42%	62.73%	61.31%	63.96%	64.78%	65.54%	38.74%	22.72%	32.27%
none	51.19%	**51.73%**	62.78%	62.27%	64.21%	65.18%	66.82%	**39.08%**	22.10%	**34.78%**

*No training data met this cut-off.

Table 4.14: Labeled attachment scores (LAS) for cross-entropy difference selection from multiple source languages

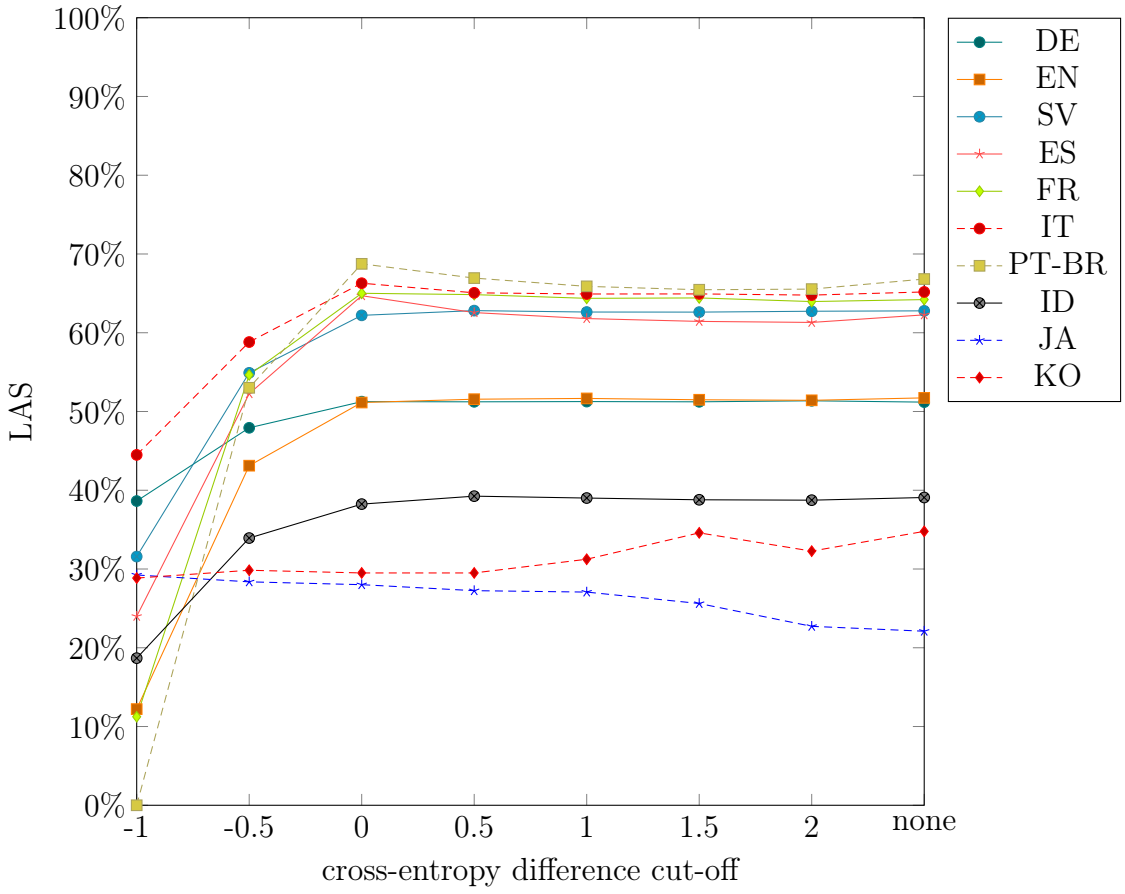

Figure 4.4: Labeled attachment scores (LAS) with different cross-entropy difference thresholds for selecting training data from multiple source languages

4.2.2 English Only as Source Language

While using multiple source languages allows for the possibility of selecting training data that is closer to the test data, a pool of training data in multiple languages may not always be available. Since English training data is much more prevalent and readily available for a variety of tasks, the instance selection experiments are repeated here using English only as the source language.

Random Selection Baseline

The random selection baseline results for selecting training data from English are shown in Tables 4.15 and 4.16. Randomly selecting as few as 1,000 sentences provides results no more than three percentage points worse than using all of the available training data for all languages tested, both for UAS and LAS. For most languages, the performance improves as more sentences are used for training. Japanese and Korean results, however, look a bit different from the other languages. For both languages, all the results for both UAS and LAS, only vary by at most 1.5%, with no clear trends in the variance.

Number of Sentences	Target Test Language								
	Unlabeled Attachment Score (UAS)								
	Germanic		Romance						
	DE	SV	ES	FR	IT	PT-BR	ID	JA	KO
1,000	56.31%	67.71%	64.09%	65.78%	65.66%	64.52%	40.97%	42.20%	41.74%
5,000	56.76%	68.35%	64.02%	67.05%	65.88%	64.57%	41.77%	42.01%	41.79%
10,000	57.19%	68.90%	64.53%	67.59%	66.59%	65.03%	41.78%	41.59%	41.62%
20,000	57.30%	69.34%	64.90%	67.88%	67.09%	65.43%	42.04%	40.87%	40.89%
all	57.15%	70.00%	65.30%	68.44%	68.00%	65.82%	42.40%	41.14%	40.97%

Table 4.15: Unlabeled attachment scores (UAS) for random selection from English (average of five trials)

Number of Sentences	Target Test Language								
	Labeled Attachment Score (LAS)								
	Germanic		Romance						
	DE	SV	ES	FR	IT	PT-BR	ID	JA	KO
1,000	48.12%	57.41%	54.23%	56.12%	56.46%	58.53%	34.99%	10.91%	25.88%
5,000	48.28%	58.04%	54.12%	57.05%	56.78%	58.52%	35.67%	10.76%	25.79%
10,000	48.64%	58.37%	54.47%	57.54%	57.34%	59.00%	35.69%	10.19%	25.66%
20,000	48.64%	58.71%	54.70%	57.86%	57.72%	59.25%	35.90%	9.88%	24.90%
all	48.54%	59.26%	55.16%	58.27%	58.55%	59.65%	36.33%	9.77%	24.72%

Table 4.16: Labeled attachment scores (LAS) for random selection from English (average of five trials)

Perplexity Selection

The results for selecting a set number of sentences from the English training data based on perplexity are shown in Tables 4.17 and 4.18. Selecting using perplexity doesn't show any improvement over random selection for most of the languages. Korean, however, has a .74% increase over the highest random selection UAS, which actually marks a 1.56% increase over using all of the training data. Similarly, Japanese has a 1.36% increase over using all the data when selecting 20,000 sentences using perplexity, but this score is only .3% increase over the highest UAS achieved using random selection. For both languages, an increase of about 1% in LAS from using all the training data is found when selecting fewer sentences using perplexity selection, but this is not a significant increase over randomly selecting fewer sentences. This contrasts strongly with the clear improvement for Japanese and Korean when using perplexity selection from multiple source languages. These results seem to indicate that these selection techniques are useful when relevant data is available but infrequent in the training data, as in the case of the pool of multiple source languages where the Japanese training data works well for parsing Korean and vice versa but must be retrieved from amongst the other languages. When there isn't relevant training data, as is the case with using only English as the training data for a language like Japanese or Korean, this selection technique is ineffective.

The results for selecting training data from English based on a perplexity threshold are shown in Tables 4.19 and 4.20 and Figures 4.5 and 4.6. Although there is still no improvement over random selection for most of the languages, using a cut-off rather than a set number of sentences provides somewhat more promising results for Japanese and Korean. In the case of Japanese, a boost of 5.05% over the highest random selection UAS and 1.81% over the highest random selection LAS is seen, whereas with Korean, a boost of 4.35% in UAS is seen, although there is no boost in LAS. Interestingly, the highest scores for UAS for both languages are achieved when

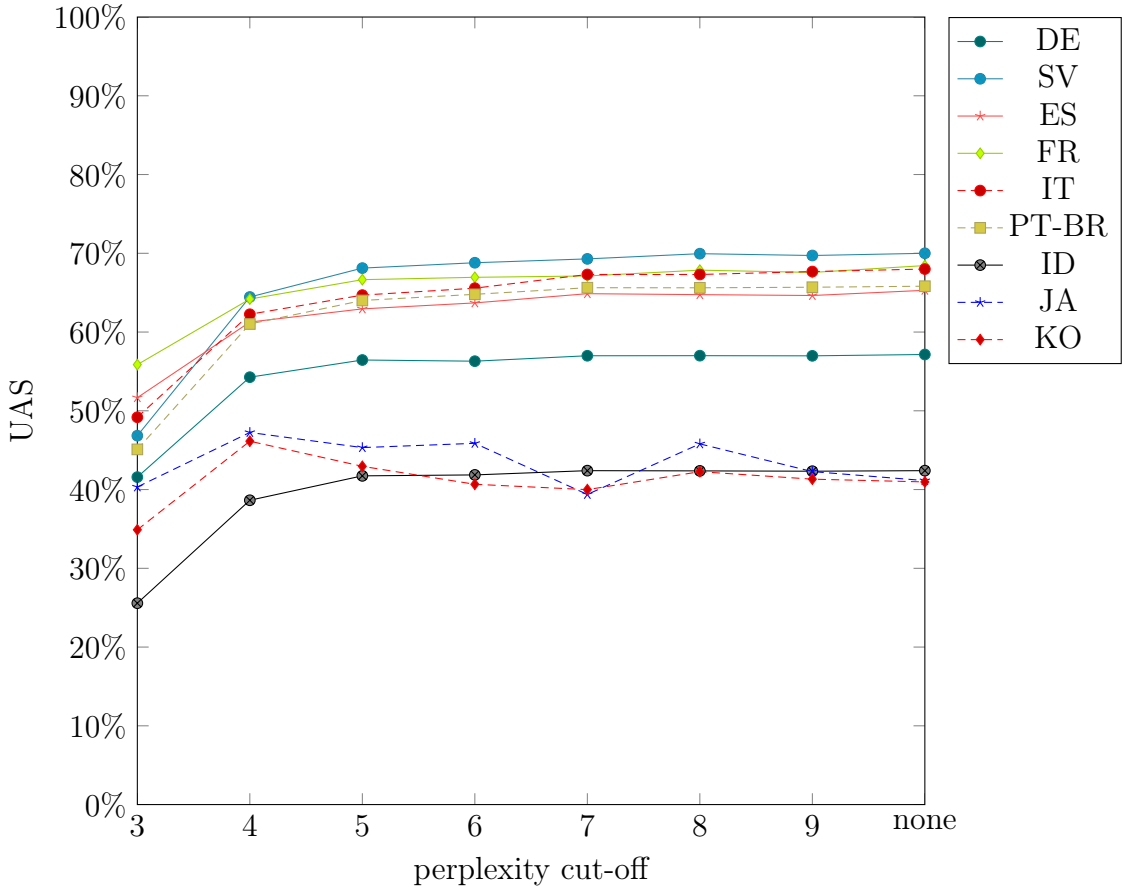

Figure 4.5: Unlabeled attachment scores (UAS) with different perplexity thresholds for selecting training data from English

using a perplexity cut-off of 4, although the highest scores for LAS occur when using a cut-off of 8. Unlike when using multiple source languages, no set perplexity threshold can be chosen that comes close to maximizing the scores for all languages tested; a cut-off of 4, although ideal for Japanese and Korean UAS, yields a significant drop in all other languages over their best possible UAS, while a higher cut-off comes closer to best possible UAS for other languages, but provides significantly worse results for Japanese and Korean.

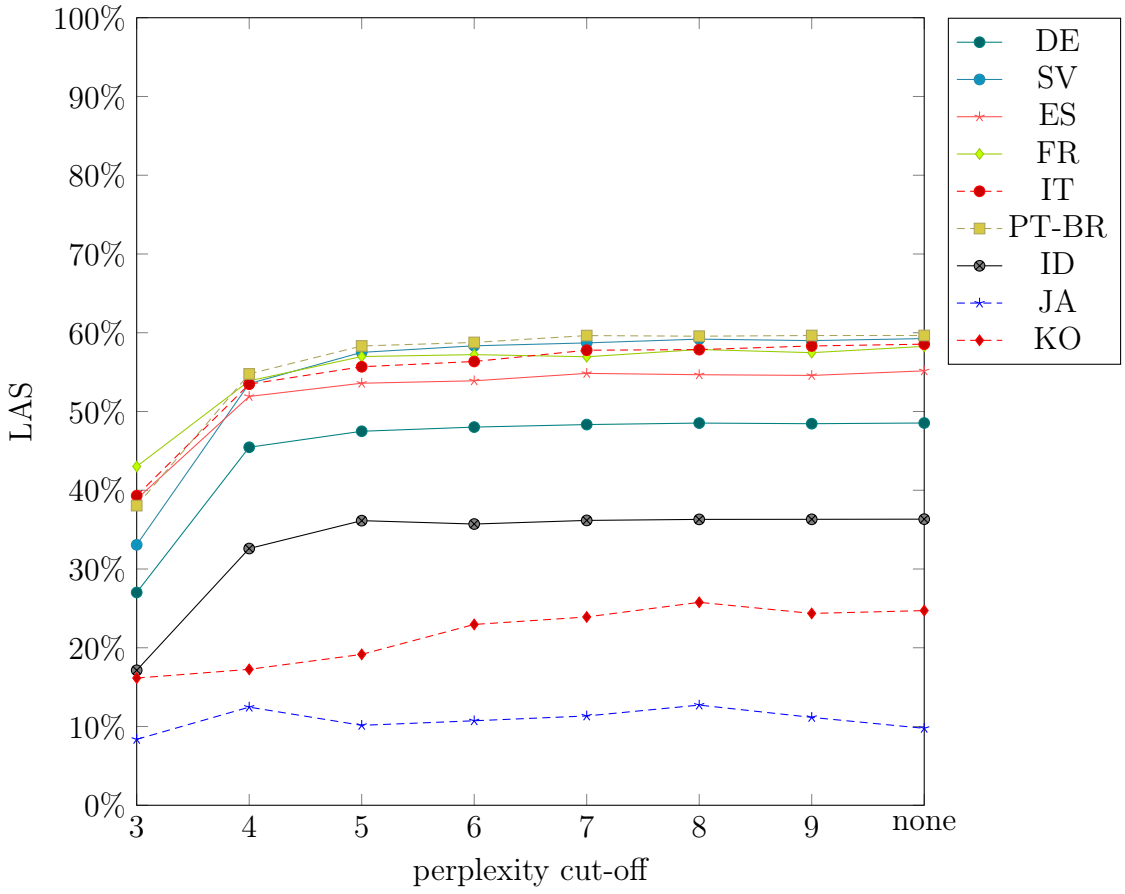

Figure 4.6: Labeled attachment scores (LAS) with different perplexity thresholds for selecting training data from English

Number of Sentences	Target Test Language								
	Unlabeled Attachment Score (UAS)								
	Germanic		Romance						
	DE	SV	ES	FR	IT	PT-BR	ID	JA	KO
1,000	56.41%	65.13%	61.94%	63.67%	65.14%	61.73%	41.10%	40.33%	41.85%
5,000	56.98%	67.88%	63.58%	67.05%	64.46%	64.84%	41.98%	41.23%	**42.53%**
10,000	57.12%	68.67%	64.75%	68.19%	65.35%	65.44%	**42.50%**	42.46%	41.43%
20,000	**57.20%**	69.69%	64.83%	67.93%	66.61%	65.57%	42.11%	**42.50%**	41.28%
all	57.15%	**70.00%**	**65.30%**	**68.44%**	**68.00%**	**65.82%**	42.40%	41.14%	40.97%

Table 4.17: Unlabeled attachment scores (UAS) for perplexity selection from English

Number of Sentences	Target Test Language								
	Labeled Attachment Score (LAS)								
	Germanic		Romance						
	DE	SV	ES	FR	IT	PT-BR	ID	JA	KO
1,000	47.13%	54.22%	52.41%	53.86%	55.80%	56.17%	35.65%	10.69%	25.31%
5,000	48.33%	57.31%	54.09%	57.24%	55.57%	59.01%	35.90%	**10.73%**	**26.11%**
10,000	48.32%	58.42%	54.78%	57.97%	56.16%	59.37%	36.30%	10.58%	25.88%
20,000	**48.64%**	59.01%	54.79%	57.80%	57.39%	59.45%	36.16%	10.62%	25.31%
all	48.54%	**59.26%**	**55.16%**	**58.27%**	**58.55%**	**59.65%**	**36.33%**	9.77%	24.72%

Table 4.18: Labeled attachment scores (LAS) for perplexity selection from English

Perplexity Cut-Off	Target Test Language								
	Unlabeled Attachment Score (UAS)								
	Germanic		Romance						
	DE	SV	ES	FR	IT	PT-BR	ID	JA	KO
3	41.58%	46.84%	51.67%	55.86%	49.18%	45.10%	25.57%	40.33%	34.89%
4	54.27%	64.45%	61.27%	64.19%	62.26%	61.02%	38.64%	**47.25%**	**46.14%**
5	56.45%	68.12%	62.94%	66.65%	64.69%	64.01%	41.74%	45.33%	42.95%
6	56.30%	68.80%	63.70%	66.95%	65.57%	64.79%	41.86%	45.87%	40.67%
7	56.99%	69.29%	64.87%	67.12%	67.31%	65.63%	**42.40%**	39.38%	39.99%
8	57.00%	69.95%	64.75%	67.86%	67.31%	65.62%	42.37%	45.80%	42.27%
9	56.98%	69.72%	64.64%	67.57%	67.68%	65.69%	42.33%	42.28%	41.32%
none	**57.15%**	**70.00%**	**65.30%**	**68.44%**	**68.00%**	**65.82%**	**42.40%**	41.14%	40.97%

Table 4.19: Unlabeled attachment scores (UAS) for perplexity selection from English

Perplexity Cut-Off	Target Test Language								
	Labeled Attachment Score (LAS)								
	Germanic		Romance						
	DE	SV	ES	FR	IT	PT-BR	ID	JA	KO
3	27.02%	33.07%	38.88%	43.02%	39.31%	38.05%	17.14%	8.37%	16.15%
4	45.46%	53.54%	51.91%	53.89%	53.47%	54.77%	32.60%	12.46%	17.26%
5	47.49%	57.51%	53.59%	56.98%	55.68%	58.32%	36.14%	10.15%	19.16%
6	48.02%	58.34%	53.90%	57.22%	56.35%	58.77%	35.71%	10.73%	22.96%
7	48.34%	58.71%	54.84%	56.95%	57.78%	59.64%	36.17%	11.34%	23.91%
8	48.53%	59.19%	54.67%	57.87%	57.86%	59.56%	36.30%	**12.72%**	**25.77%**
9	48.45%	59.00%	54.58%	57.47%	58.32%	**59.65%**	36.31%	11.16%	24.36%
none	**48.54%**	**59.26%**	**55.16%**	**58.27%**	**58.55%**	**59.65%**	**36.33%**	9.77%	24.72%

Table 4.20: Labeled attachment scores (LAS) for perplexity selection from English

Cross-Entropy Difference Selection

The results for selecting a set number of sentences from the training data based on cross-entropy difference are shown in Tables 4.21 and 4.22. While these results don't show a huge improvement over random selection, there are some interesting trends. For all languages except German and Italian, the highest UAS occurs when fewer sentences are selected, and this score is consistently around a percentage point higher than selecting the same number of sentences randomly. However, for all but Japanese and Korean, this score is only at most around half a percentage point higher than using all of the English training data. For Japanese and Korean, it is around 1.5% higher than using all the data, but for Japanese, it isn't significantly higher than the highest score achieved using random selection, and for Korean, it is less than a percentage point higher than the highest score from random selection.

The results for LAS are not as promising, although some languages show a similar trend as with UAS. French and Brazilian Portuguese achieve scores very slightly higher than using all of the training data and close to but not quite a percentage point higher than randomly selecting the same number of sentences. Indonesian achieves a highest LAS about half a percentage point higher than using all the training data and over a percentage point higher than randomly selecting the same number of sentences. The highest LAS for Japanese and Korean occur when selecting fewer sentences but aren't significantly higher than randomly selecting the same number of sentences.

The results for selecting from the training data based on a cross-entropy difference threshold are shown in Tables 4.23 and 4.24 and Figures 4.7 and 4.8. Similar to using a perplexity threshold for selecting from English training data, Japanese and Korean achieve significantly higher scores (more than 5% UAS increase and around 4% LAS increase over using all the training data for both languages), with a relatively low cut-off. Once again, there is no clear threshold that comes close to maximizing scores across languages.

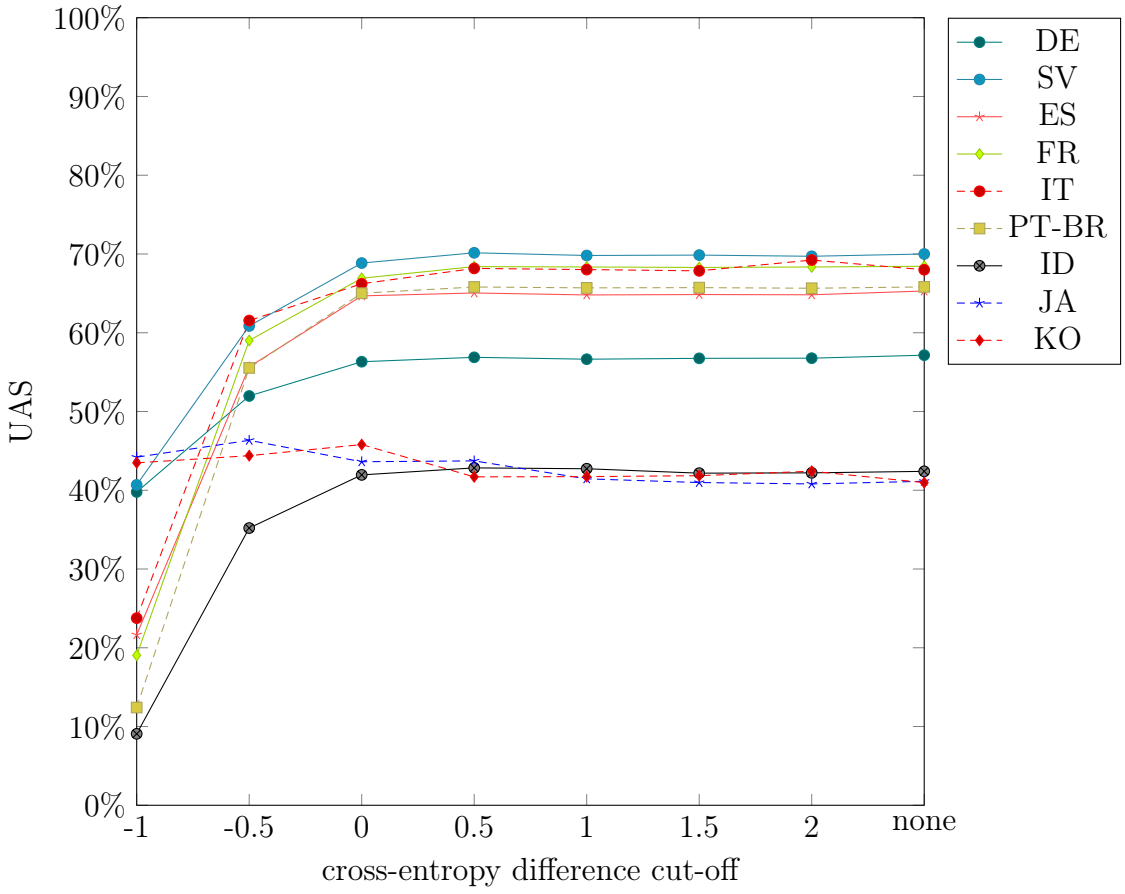

Figure 4.7: Unlabeled attachment scores (UAS) with different cross-entropy difference thresholds for selecting training data from English

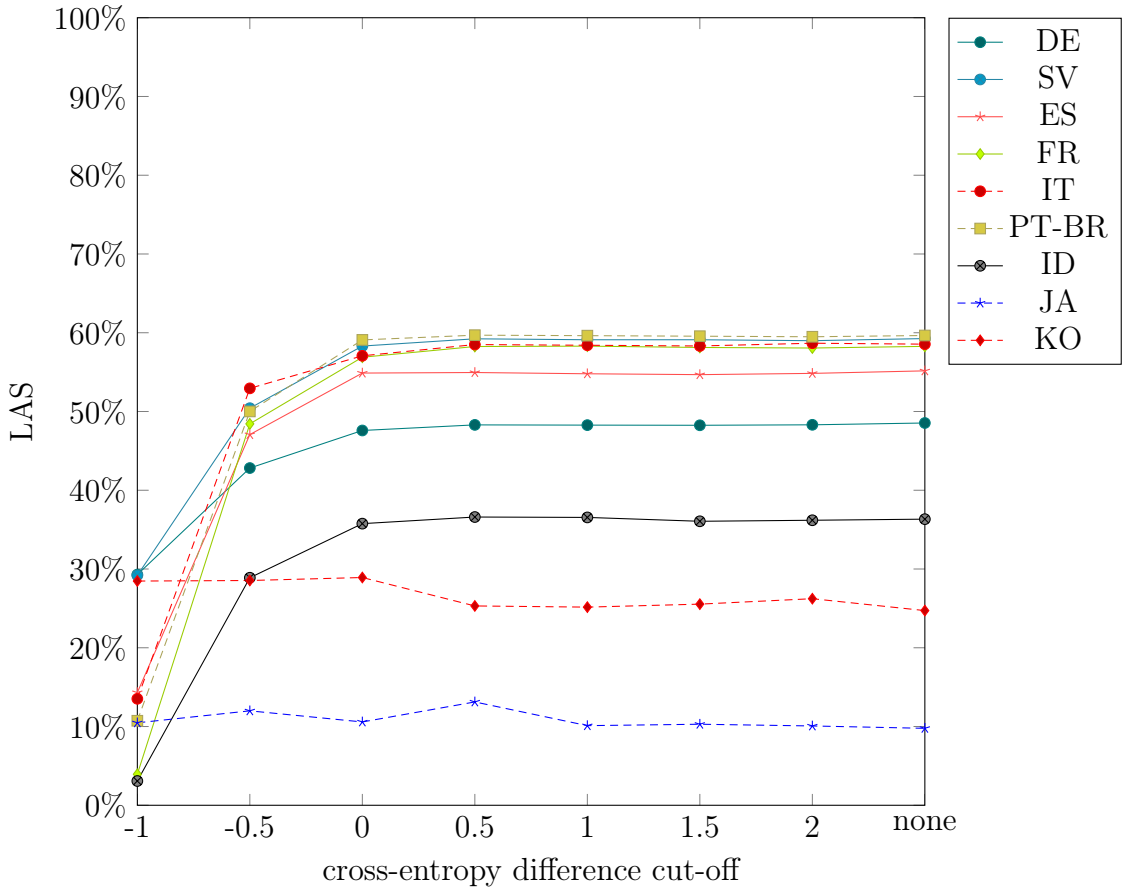

Figure 4.8: Labeled attachment scores (LAS) with different cross-entropy difference thresholds for selecting training data from English

Number of Sentences	Target Test Language								
	Unlabeled Attachment Score (UAS)								
	Germanic		Romance						
	DE	SV	ES	FR	IT	PT-BR	ID	JA	KO
1,000	55.52%	67.24%	63.54%	67.48%	65.39%	64.35%	41.64%	40.11%	**42.61%**
5,000	56.57%	69.30%	65.41%	67.34%	67.29%	65.16%	42.23%	42.25%	42.46%
10,000	56.82%	69.33%	**65.56%**	68.07%	67.30%	**66.07%**	**42.95%**	42.07%	41.62%
20,000	56.86%	**70.06%**	64.88%	**68.55%**	67.37%	65.51%	42.83%	**42.36%**	41.09%
all	**57.15%**	70.00%	65.30%	68.44%	**68.00%**	65.82%	42.40%	41.14%	40.97%

Table 4.21: Unlabeled attachment scores (UAS) for cross-entropy difference selection from English

Number of Sentences	Target Test Language								
	Labeled Attachment Score (LAS)								
	Germanic		Romance						
	DE	SV	ES	FR	IT	PT-BR	ID	JA	KO
1,000	47.15%	56.47%	54.25%	57.30%	56.13%	58.47%	35.78%	10.33%	**25.81%**
5,000	47.92%	58.73%	55.05%	57.05%	57.91%	59.17%	36.20%	**10.83%**	25.43%
10,000	48.23%	58.81%	55.12%	58.06%	57.90%	**59.88%**	**36.81%**	10.40%	25.05%
20,000	48.46%	59.06%	54.74%	**58.46%**	57.93%	59.36%	36.61%	10.36%	25.05%
all	**48.54%**	**59.26%**	**55.16%**	58.27%	**58.55%**	59.65%	36.33%	9.77%	24.72%

Table 4.22: Labeled attachment scores (LAS) for cross-entropy difference selection from English

Cross-Entropy Difference Cut-Off	Target Test Language								
	Unlabeled Attachment Score (UAS)								
	Germanic		Romance						
	DE	SV	ES	FR	IT	PT-BR	ID	JA	KO
-1	39.78%	40.68%	21.65%	19.04%	23.76%	12.41%	9.07%	44.20%	43.48%
-0.5	51.97%	60.86%	55.67%	59.02%	61.55%	55.53%	35.20%	**46.34%**	44.39%
0	56.32%	68.85%	64.68%	66.92%	66.22%	65.02%	41.95%	43.62%	**45.80%**
0.5	56.88%	**70.15%**	65.05%	68.40%	68.17%	65.81%	**42.84%**	43.73%	41.70%
1	56.64%	69.81%	64.80%	68.35%	68.02%	65.70%	42.73%	41.45%	41.73%
1.5	56.75%	69.86%	64.85%	68.29%	67.86%	65.73%	42.17%	40.98%	41.85%
2	56.77%	69.71%	64.82%	68.35%	**69.23%**	65.65%	42.21%	40.80%	42.42%
none	**57.15%**	70.00%	**65.30%**	**68.44%**	68.00%	**65.82%**	42.40%	41.14%	40.97%

Table 4.23: Unlabeled attachment scores (UAS) for cross-entropy difference selection from English

Cross-Entropy Difference Cut-Off	Target Test Language								
	Unlabeled Attachment Score (UAS)								
	Germanic		Romance						
	DE	SV	ES	FR	IT	PT-BR	ID	JA	KO
-1	29.29%	29.19%	14.31%	3.94%	13.52%	10.73%	3.08%	10.47%	28.47%
-0.5	42.82%	50.42%	47.07%	48.43%	52.94%	50.01%	28.89%	11.99%	28.54%
0	47.59%	58.31%	54.88%	56.89%	57.07%	59.09%	35.76%	10.58%	**28.92%**
0.5	48.30%	59.23%	54.95%	58.25%	58.50%	**59.69%**	**36.60%**	**13.12%**	25.31%
1	48.27%	59.11%	54.79%	**58.30%**	58.41%	59.63%	36.55%	10.11%	25.16%
1.5	48.25%	59.11%	54.68%	58.13%	58.32%	59.56%	36.06%	10.29%	25.54%
2	48.31%	58.99%	54.84%	58.06%	**58.67%**	59.49%	36.19%	10.07%	26.23%
none	**48.54%**	**59.26%**	**55.16%**	58.27%	58.55%	59.65%	36.33%	9.77%	24.72%

Table 4.24: Labeled attachment scores (LAS) for cross-entropy difference selection from English

4.2.3 Rearranged English as Source Language

In order to leverage language-specific knowledge, this set of experiments uses rearranged English as a source language. All occurrences of adjectives preceding the noun they modify are rearranged so that the adjectives occur after the noun, while maintaining all other ordering and dependencies. This rearranged English is then used as the source language for the three target languages which have the most occurrences of modifying adjectives following the noun relative to modifying adjectives preceding the noun; these are Indonesian, Spanish, and French, where over 70% of immediately adjacent modifying adjectives occur following the noun rather than preceding it (in the case of Indonesian, this is actually over 90%).

Table 4.25 shows the frequencies of adjectives and nouns in the training data. Nouns are very frequent in all languages, accounting for at least a fourth of all tokens and, in the case of Indonesian, Japanese, and Korean, close to half of all tokens. Adjectives are less frequent, but still account for between about 3.5% and 8.5% of tokens. This indicates that correcting errors in adjective dependencies could provide a significant boost to scores and that nouns may be a good way to do this since they're so frequent and thus could prove a reliable indicator.

Table 4.26 shows the position of immediately adjacent modifying adjectives. The counts shown are the number of occurrences where the adjective is immediately adjacent to the noun it modifies, either preceding or following the noun. These counts make it clear that using English training data, where immediately adjacent modifying adjectives overwhelmingly (97.93% of the time) precede the noun they modify, is likely to cause errors where parsing languages like Indonesian where immediately adjacent modifying adjectives overwhelmingly (91.31% of the time) follow the noun they modify. This is likely to also be the case for languages where there is still a significant proportion of immediately adjacent modifying adjectives that precede the noun, but where it is significantly more likely for them to follow the noun, as is

the case with Spanish and French, where immediately adjacent modifying adjectives follow the noun 73.04% and 71.11% of the time, respectively.

language	# of sentences in training set	# of tokens in training set	# of nouns	% of tokens	# of adjectives	% of tokens
DE	14,118	264,906	78,736	29.72%	19,450	7.34%
EN	39,832	950,028	286,930	30.20%	66,402	6.99%
SV	4,447	66,631	17,095	25.66%	5,694	8.55%
ES	14,138	375,180	105,282	28.06%	21,228	5.66%
FR	14,511	351,233	93,565	26.64%	19,928	5.67%
IT	6,389	149,145	38,574	25.86%	10,587	7.10%
PT-BR	9,600	239,012	72,259	30.23%	12,138	5.08%
ID	4,477	97,531	40,238	41.26%	3,649	3.74%
JA	8,277	80,172	43,926	54.79%	2,855	3.56%
KO	5,437	60,621	28,009	46.20%	2,312	3.81%

Table 4.25: Adjective and noun frequencies

Results for doing this rearrangement and training on the full English training dataset are shown in Tables 4.27 and 4.28. Although this modified dataset still has worse scores than using multiple source languages for all three test languages and results in a slight drop in scores for Spanish over non-rearranged English, there is a

boost over non-rearranged English for French and Indonesian. For French, this boost is about half a percentage point for both UAS and LAS, while for Indonesian, it is about a full percentage point. It seems reasonable that Indonesian would have the biggest boost from this technique, since it overwhelmingly has the highest relative frequency of modifying adjectives following the nouns they modify.

language	# of mod. adj. preceding noun	% of all adj.	% of adjacent mod. adj.	# of mod. adj. following noun	% of all adj.	% of adjacent mod. adj.
DE	11,197	57.57%	99.64%	41	0.21%	0.36%
EN	37,602	56.63%	97.93%	793	1.19%	2.07%
SV	3,561	62.54%	98.89%	40	0.70%	1.11%
ES	4,099	19.31%	26.96%	11,105	52.31%	73.04%
FR	4,179	20.97%	28.89%	10,288	51.63%	71.11%
IT	2,934	27.71%	37.89%	4,810	45.43%	62.11%
PT-BR	2,709	22.32%	31.54%	5,879	48.43%	68.46%
ID	160	4.38%	8.69%	1,682	46.09%	91.31%
JA	1,535	53.77%	100.00%	0	0.00%	0.00%
KO	761	32.92%	100.00%	0	0.00%	0.00%

Table 4.26: Position of modifying adjectives

Training Dataset	Target Test Language		
	Unlabeled Attachment Score (UAS)		
	Romance		
	ES	FR	ID
same language (upperbound)	76.48%	77.27%	77.76%
multiple source languages	71.90%	73.81%	45.64%
English	65.30%	68.44%	42.40%
rearranged English	64.73%	69.02%	43.61%

Table 4.27: Unlabeled attachment scores (UAS) for full training datasets

Training Dataset	Target Test Language		
	Labeled Attachment Score (LAS)		
	Romance		
	ES	FR	ID
same language (upperbound)	68.45%	69.54%	70.86%
multiple source languages	62.27%	64.21%	39.08%
English	55.16%	58.27%	36.33%
rearranged English	54.47%	58.68%	37.30%

Table 4.28: Labeled attachment scores (LAS) for full training datasets

Random Selection Baseline

The random selection baseline results for selecting training data from rearranged English are shown in Tables 4.29 and 4.30. When comparing random selection of a set number of sentences from English and from rearranged English, the trends seen when using the full training dataset remain consistent. That is, Spanish scores are consistently around the same or slightly lower, French scores are consistently half a percentage point higher, and Indonesian scores are consistently a percentage point higher.

Number of Sentences	**Target Test Language**		
	Unlabeled Attachment Score (UAS)		
	Romance		
	ES	**FR**	**ID**
1,000	64.48%	67.59%	42.32%
5,000	64.40%	68.48%	42.99%
10,000	64.76%	69.04%	43.19%
20,000	64.64%	68.67%	43.65%
all	64.73%	69.02%	43.61%

Table 4.29: Unlabeled attachment scores (UAS) for random selection from rearranged English (average of five trials)

Number of Sentences	Target Test Language		
	Labeled Attachment Score (LAS)		
	Romance		
	ES	FR	ID
1,000	54.58%	57.60%	36.30%
5,000	54.39%	58.48%	37.03%
10,000	54.69%	58.87%	36.98%
20,000	54.43%	58.60%	37.39%
all	54.47%	58.68%	37.30%

Table 4.30: Labeled attachment scores (LAS) for random selection from rearranged English (average of five trials)

Perplexity Selection

The results for selecting a set number of sentences from the rearranged English training data based on perplexity are shown in Tables 4.31 and 4.32. For Spanish and French, selection of 20,000 sentences based on perplexity provides approximately half a percentage point boost in UAS and LAS compared to using all the training data. For Indonesian, however, no boost in UAS is seen, although selecting 10,000 sentences gives a LAS slightly higher than using all the training data and around half a percentage point higher than randomly selecting 10,000 sentences.

The results for selecting training data from rearranged English based on a perplexity threshold are shown in Tables 4.33 and 4.34 and Figures 4.9 and 4.10. These results look similar to selecting a set number of sentences; the highest scores achieved

Number of Sentences	Target Test Language		
	Unlabeled Attachment Score (UAS)		
	Romance		
	ES	FR	ID
1,000	62.77%	67.02%	42.80%
5,000	64.24%	68.71%	42.92%
10,000	65.00%	69.12%	43.44%
20,000	**65.26%**	**69.57%**	43.52%
all	64.73%	69.02%	**43.61%**

Table 4.31: Unlabeled attachment scores (UAS) for perplexity selection from rearranged English

with a threshold are not significantly different from the highest scores achieved by selecting a set number of sentences. Using a cut-off of 8 for perplexity gives the best UAS for French and best LAS for Spanish and Indonesian, while coming within .2% of the best UAS for Spanish and Indonesian and the best LAS for French.

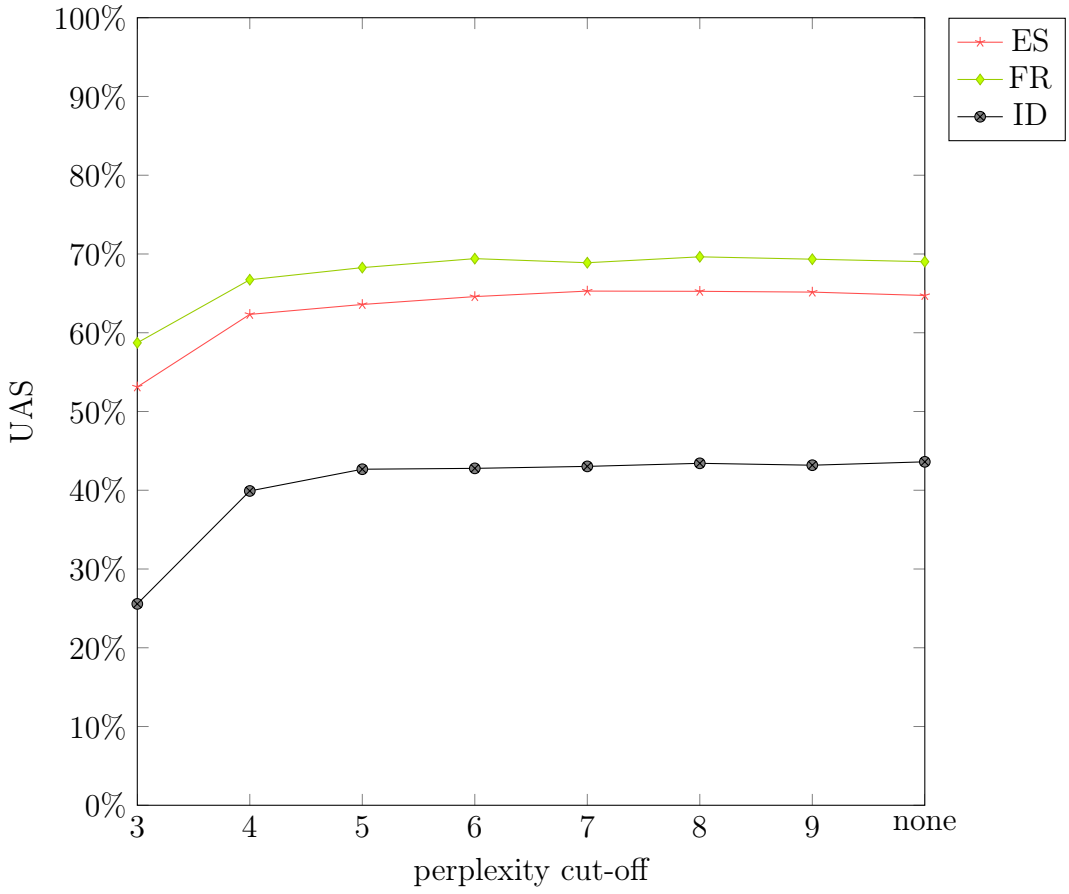

Figure 4.9: Unlabeled attachment scores (UAS) with different perplexity thresholds for selecting training data from rearranged English

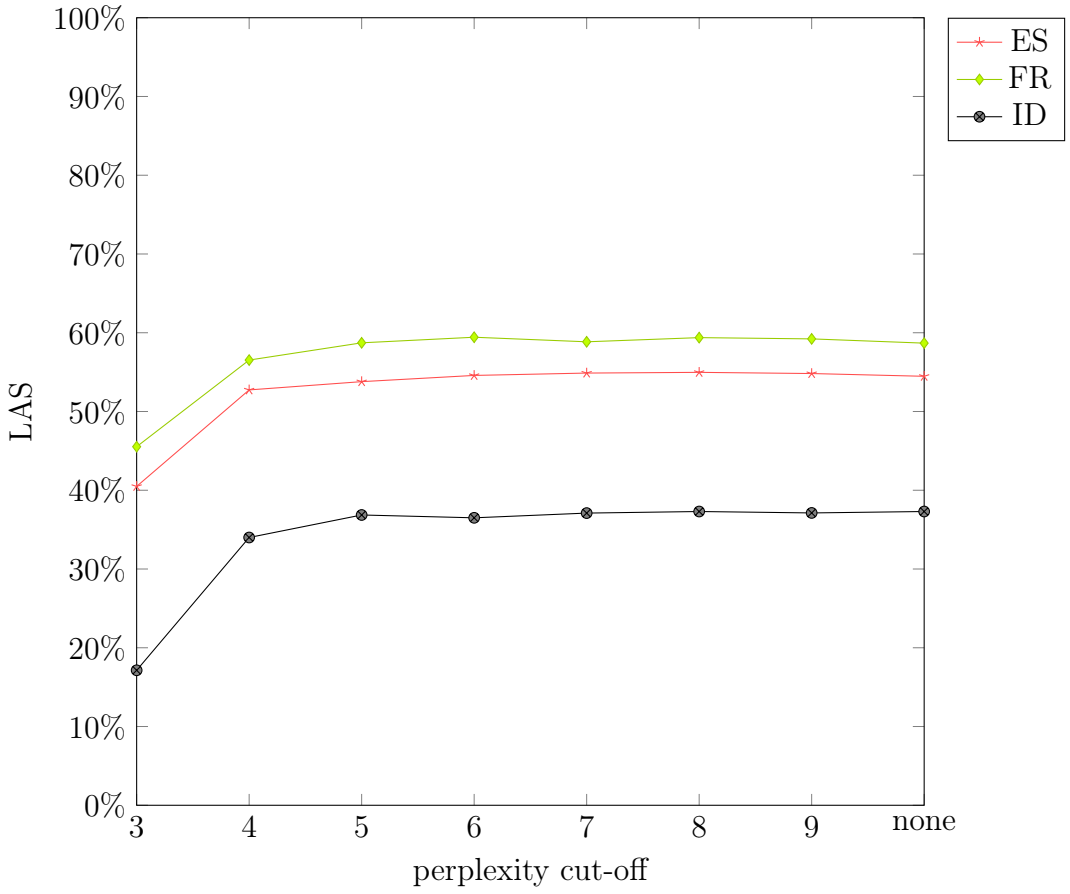

Figure 4.10: Labeled attachment scores (LAS) with different perplexity thresholds for selecting training data from rearranged English

Number of Sentences	Target Test Language		
	Labeled Attachment Score (LAS)		
	Romance		
	ES	FR	ID
1,000	52.97%	56.65%	36.67%
5,000	54.43%	58.81%	36.80%
10,000	54.73%	59.04%	**37.42%**
20,000	**54.95%**	**59.48%**	37.41%
all	54.47%	58.68%	37.30%

Table 4.32: Labeled attachment scores (LAS) for perplexity selection from rearranged English

Perplexity Cut-Off	Target Test Language		
	Unlabeled Attachment Score (UAS)		
	Romance		
	ES	FR	ID
3	53.12%	58.72%	25.57%
4	62.33%	66.73%	39.91%
5	63.59%	68.27%	42.67%
6	64.59%	69.41%	42.78%
7	**65.29%**	68.89%	43.03%
8	65.26%	**69.64%**	43.42%
9	65.16%	69.34%	43.18%
none	64.73%	69.02%	**43.61%**

Table 4.33: Unlabeled attachment scores (UAS) for perplexity selection from rearranged English

Perplexity Cut-Off	Target Test Language		
	Labeled Attachment Score (LAS)		
	Romance		
	ES	FR	ID
3	40.49%	45.53%	17.14%
4	52.74%	56.53%	34.00%
5	53.79%	58.72%	36.86%
6	54.58%	**59.43%**	36.50%
7	54.88%	58.85%	37.10%
8	**54.97%**	59.38%	**37.30%**
9	54.82%	59.22%	37.12%
none	54.47%	58.68%	**37.30%**

Table 4.34: Labeled attachment scores (LAS) for perplexity selection from rearranged English

Cross-Entropy Difference Selection

The results for selecting a set number of sentences from the training data based on cross-entropy difference are shown in Tables 4.35 and 4.36. For Spanish, selecting 10,000 sentences based on cross-entropy difference gives around a 1% boost in UAS and LAS compared to using all the training data or randomly selecting 10,000 sentences. For Indonesian, selecting 10,000 sentences gives results around 1% higher compared to randomly selecting 10,000 sentences, but not significantly better than using all the training data. For French, however, there is no improvement over random selection.

Number of Sentences	**Target Test Language**		
	Unlabeled Attachment Score (UAS)		
	Romance		
	ES	**FR**	**ID**
1,000	64.06%	66.50%	42.79%
5,000	64.30%	67.64%	43.70%
10,000	**65.53%**	68.23%	**43.89%**
20,000	65.12%	68.58%	43.87%
all	64.73%	**69.02%**	43.61%

Table 4.35: Unlabeled attachment scores (UAS) for cross-entropy difference selection from rearranged English

The results for selecting from the training data based on a cross-entropy difference threshold are shown in Tables 4.37 and 4.38 and Figures 4.11 and 4.12. There is no

Number of Sentences	Target Test Language		
	Labeled Attachment Score (LAS)		
	Romance		
	ES	FR	ID
1,000	54.29%	56.36%	36.82%
5,000	54.52%	57.64%	37.50%
10,000	**55.46%**	58.16%	**37.62%**
20,000	55.02%	**58.73%**	37.51%
all	54.47%	58.68%	37.30%

Table 4.36: Labeled attachment scores (LAS) for cross-entropy difference selection from rearranged English

significant boost with using a cross-entropy difference threshold. However, using a threshold of 0.5 for all languages does yield around a half a percentage point boost for Spanish UAS and LAS while no significant decrease in scores for French or Indonesian.

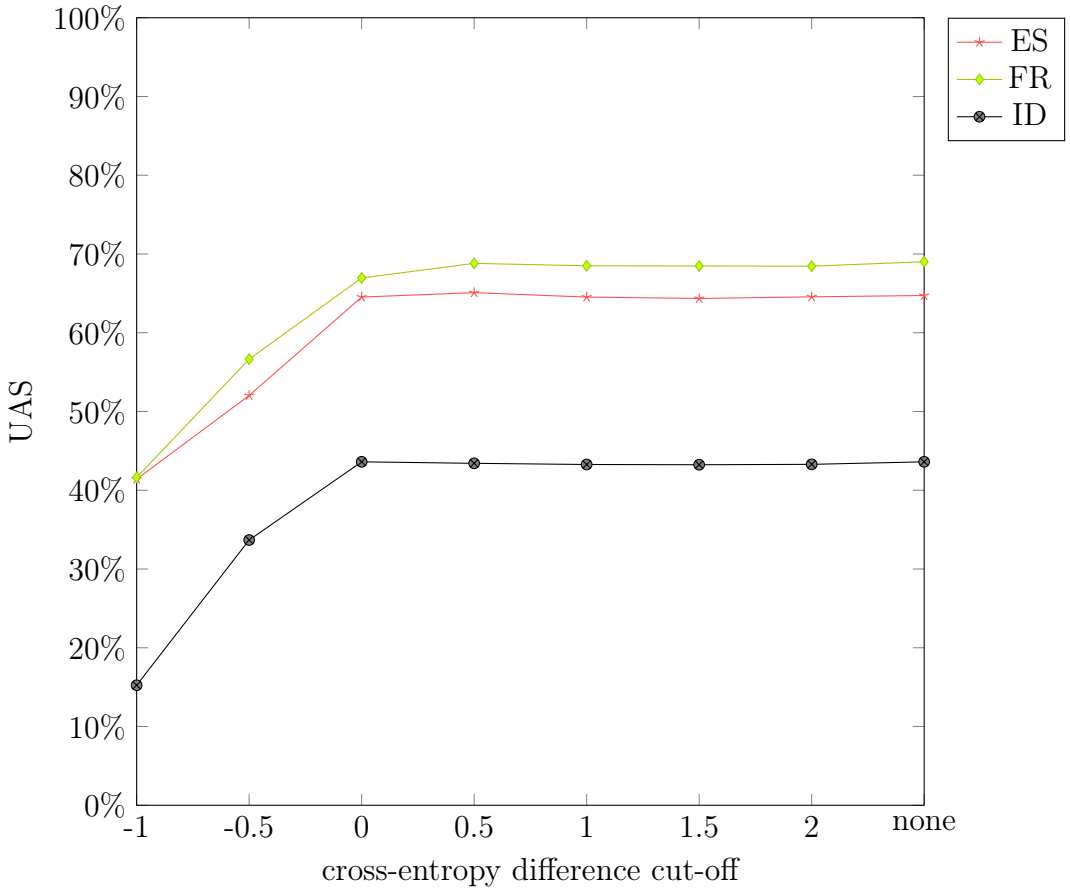

Figure 4.11: Unlabeled attachment scores (UAS) with different cross-entropy difference thresholds for selecting training data from rearranged English

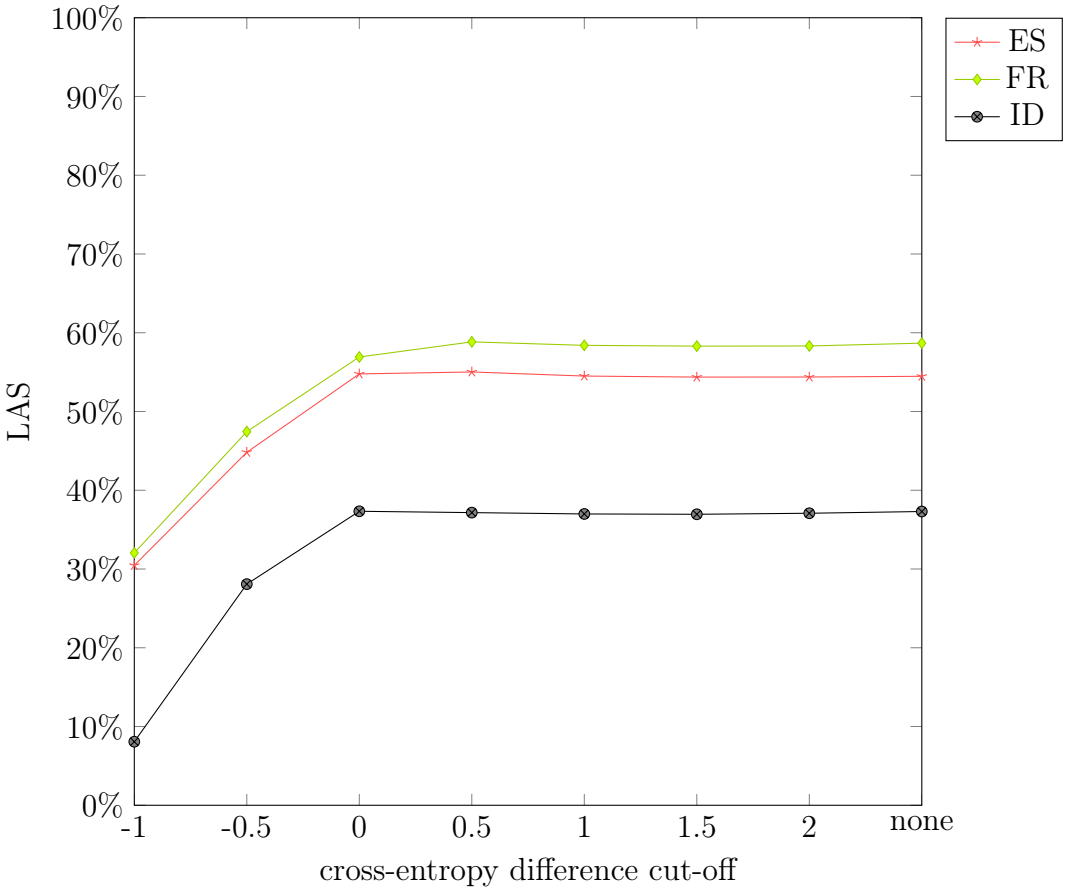

Figure 4.12: Labeled attachment scores (LAS) with different cross-entropy difference thresholds for selecting training data from rearranged English

Number of Sentences	Target Test Language		
	Unlabeled Attachment Score (UAS)		
	Romance		
	ES	FR	ID
-1	41.41%	41.63%	15.24%
-0.5	52.03%	56.65%	33.68%
0	64.53%	66.95%	**43.61%**
0.5	**65.09%**	68.81%	43.42%
1	64.53%	68.50%	43.27%
1.5	64.35%	68.48%	43.24%
2	64.56%	68.45%	43.29%
none	64.73%	**69.02%**	**43.61%**

Table 4.37: Unlabeled attachment scores (UAS) for cross-entropy difference selection from rearranged English

Number of Sentences	Target Test Language		
	Labeled Attachment Score (LAS)		
	Romance		
	ES	**FR**	**ID**
-1	30.46%	32.04%	8.07%
-0.5	44.82%	47.44%	28.08%
0	54.77%	56.92%	**37.33%**
0.5	**55.02%**	**58.84%**	37.16%
1	54.50%	58.40%	36.99%
1.5	54.37%	58.30%	36.95%
2	54.38%	58.32%	37.08%
none	54.47%	58.68%	37.30%

Table 4.38: Labeled attachment scores (LAS) for cross-entropy difference selection from rearranged English

Chapter 5

ANALYSIS

Although most of the results obtained by the methods tested are negative, some promising trends emerge. Namely, perplexity and cross-entropy difference selection seem to provide a significant boost to both unlabeled attachment scores (UAS) and labeled attachment scores (LAS) when used on a pool of multiple source languages with a target language where relevant training data are available but infrequent in the training data, with cross-entropy difference providing very slightly better performance over perplexity selection. In the case of target languages with lots of relevant training data available among multiple source languages, these selection methods don't provide the same large improvements, but also have no negative impact. However, when only English is used for training data, the results are much less clear for instance selection, and while some boost is seen for some languages, the impact is minimal. Rearranging the order of modifying adjectives relative to the nouns they modify in English training data to better match the word order seen in the target languages had a small positive impact on the results, which is promising for more extensive rearrangement. Instance selection methods applied to this rearranged data did not perform any better than on the non-rearranged data.

Figures 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11, 5.12, 5.13, 5.14, 5.15, 5.16, 5.17, and 5.18 show a synopsis of selecting a set number of sentences using three different methods - random selection, perplexity selection, and cross-entropy difference selection - and three different pools of training data - multiple source languages, English, and rearranged English, where applicable. These charts make clear that neither of the methods tested yield a consistent improvement over random, although

certain languages see an improvement when multiple source languages are used as the training data pool for selection.

Tables 5.1 and 5.2 show scores for the different instance selection methods when selecting from multiple source languages, averaged across the target languages. These scores show the general trend, where on average, higher scores are achieved with perplexity selection than with random selection, while even higher scores are achieved with cross-entropy difference selection than with perplexity selection.

Number of Sentences	**Instance Selection Method**		
	Average Unlabeled Attachment Score (UAS)		
	random selection	**perplexity selection**	**cross-entropy difference selection**
1,000	60.84%	61.14%	63.55%
5,000	62.53%	64.69%	65.75%
10,000	63.31%	64.76%	65.53%
20,000	63.81%	64.29%	64.52%
all	64.76%	64.76%	64.76%
upperbound	78.22%	78.22%	78.22%

Table 5.1: Unlabeled attachment scores (UAS) for different instance selection methods from multiple source languages averaged across all target languages

Number of Sentences	Instance Selection Method		
	Average Labeled Attachment Score (LAS)		
	random selection	**perplexity selection**	**cross-entropy difference selection**
1,000	48.55%	46.42%	49.14%
5,000	50.04%	50.16%	51.52%
10,000	50.86%	50.78%	51.80%
20,000	51.24%	51.75%	52.51%
all	52.01%	52.01%	52.01%
upperbound	69.79%	69.79%	69.79%

Table 5.2: Labeled attachment scores (LAS) for different instance selection methods from multiple source languages averaged across all target languages

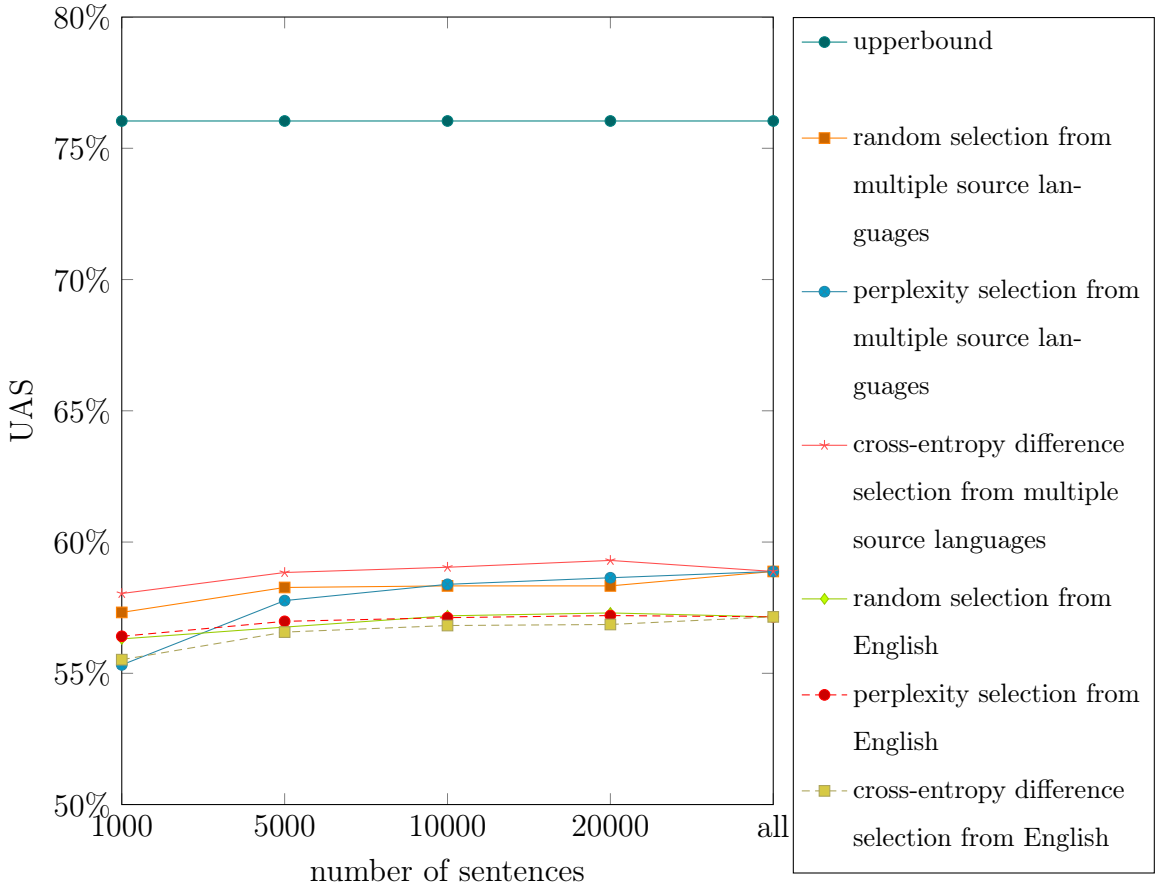

Figure 5.1: Unlabeled attachment scores (UAS) for German with different instance selection methods

5.1 Perplexity Selection

The results for perplexity selection verify and extend those shown in Søgaard (2011). When multiple source languages are used, this method of selection is a robust one for identifying relevant instances in the training data, when such instances do exist. A clear threshold emerged as a reasonable cut-off for selecting based on perplexity; with a cut-off of 6 for perplexity, the UAS and LAS for all languages (with the exception of Korean LAS) came within a percentage point of the highest seen across all cut-offs and number of sentences selected (including using all the data). Interestingly, but

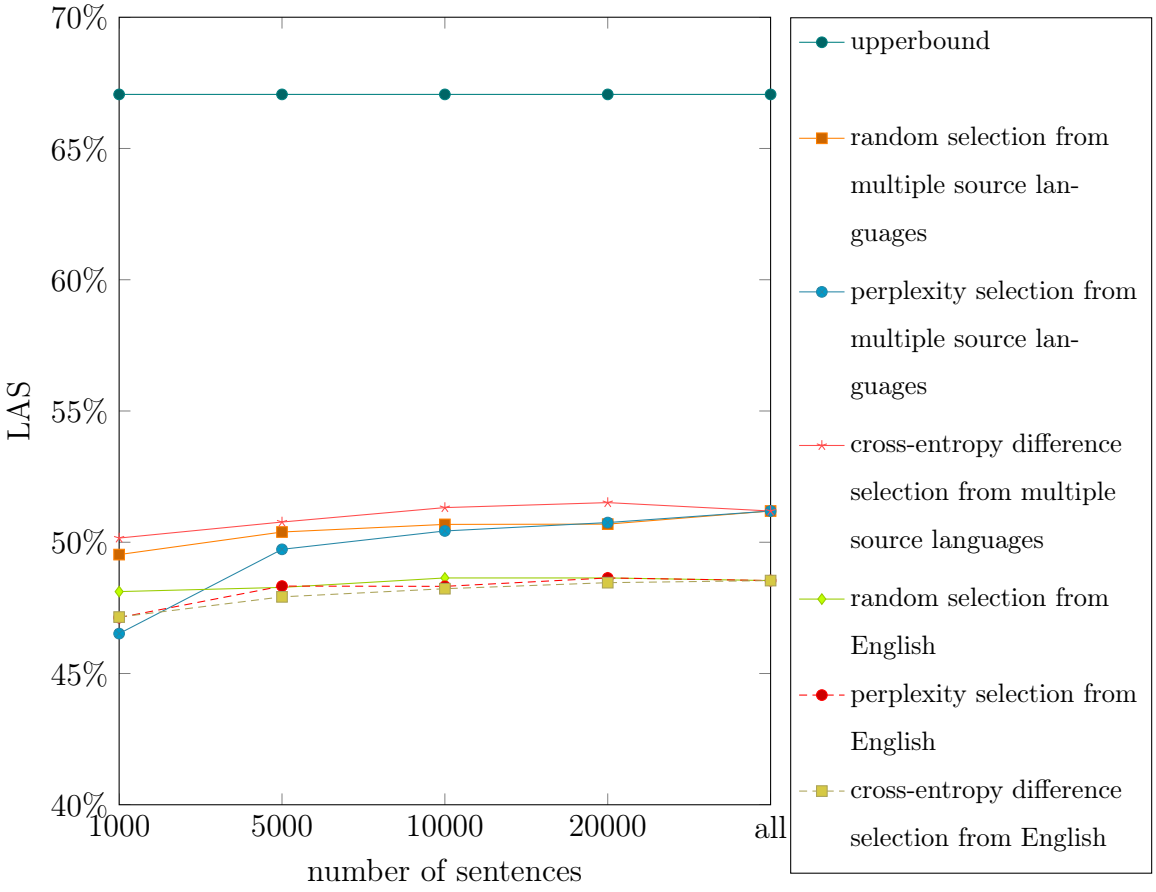

Figure 5.2: Labeled attachment scores (LAS) for German with different instance selection methods

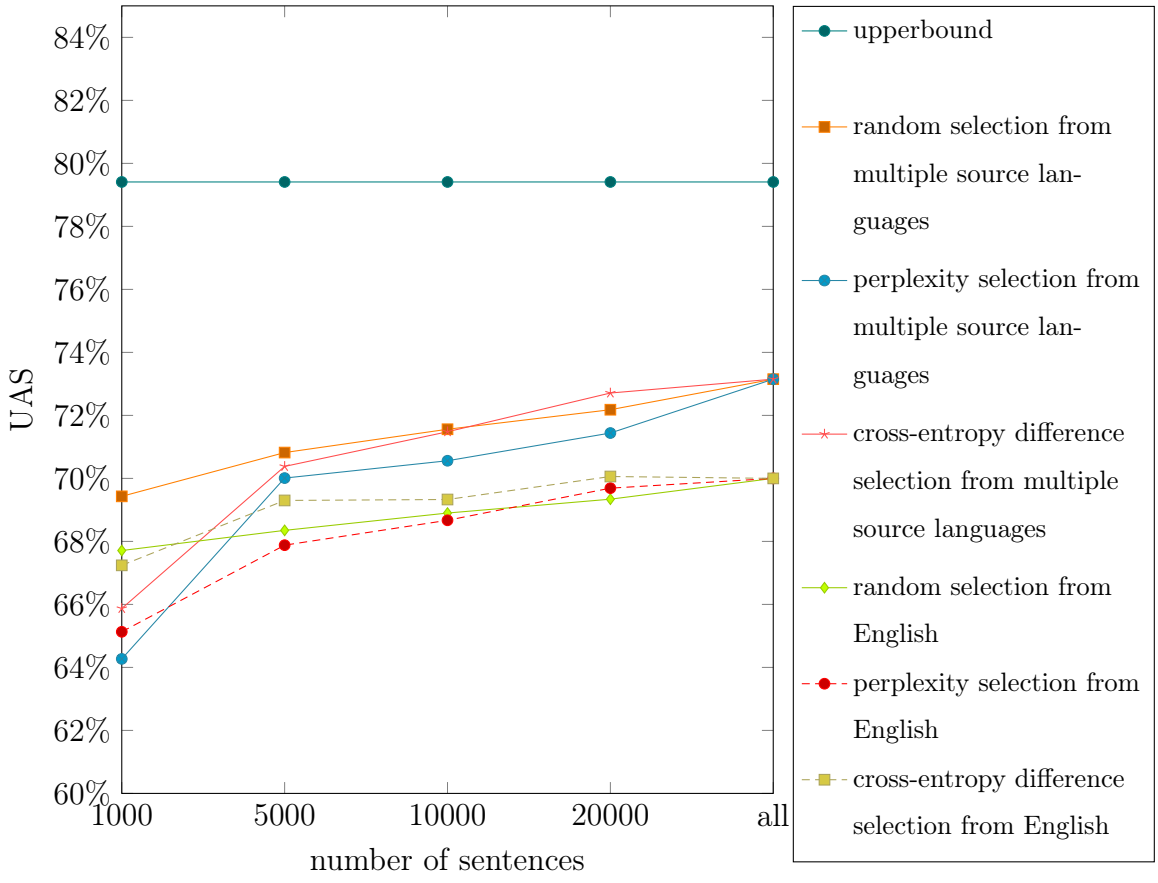

Figure 5.3: Unlabeled attachment scores (UAS) for Swedish with different instance selection methods

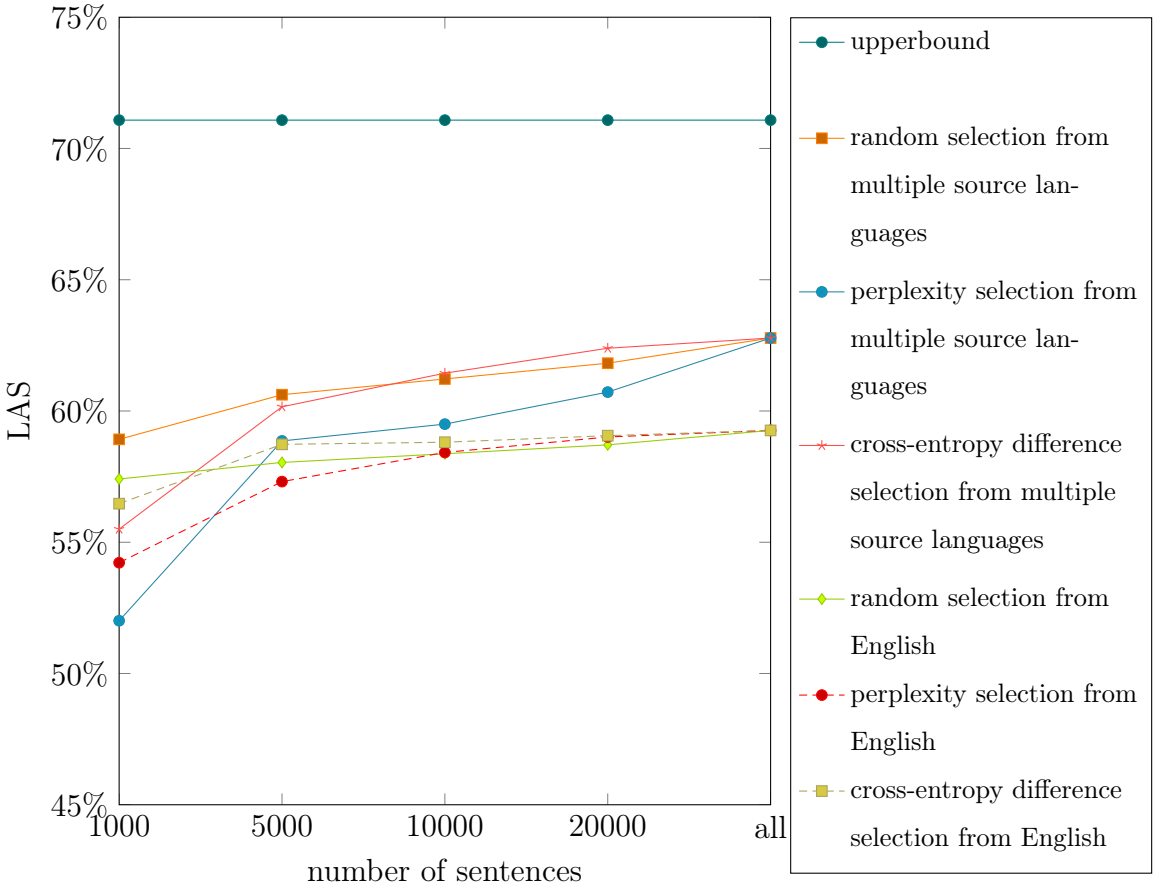

Figure 5.4: Labeled attachment scores (LAS) for Swedish with different instance selection methods

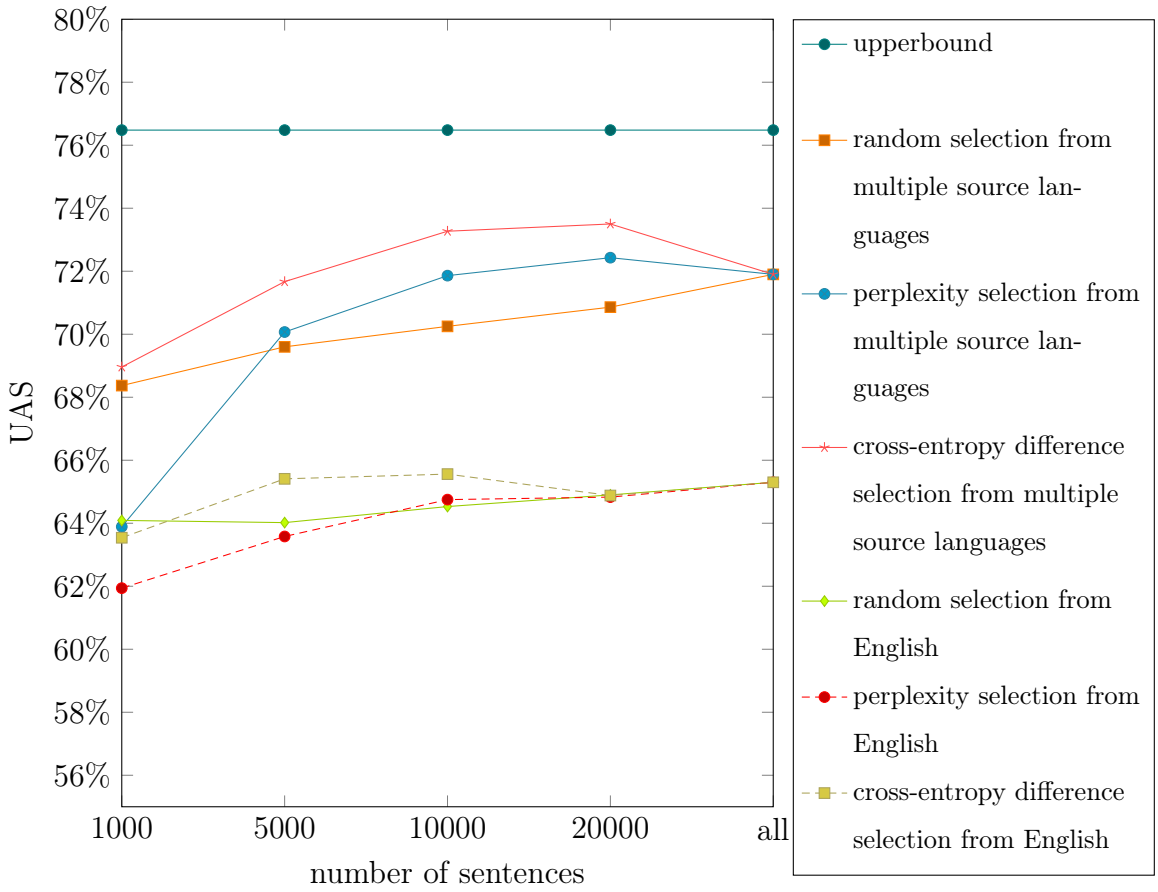

Figure 5.5: Unlabeled attachment scores (UAS) for Spanish with different instance selection methods

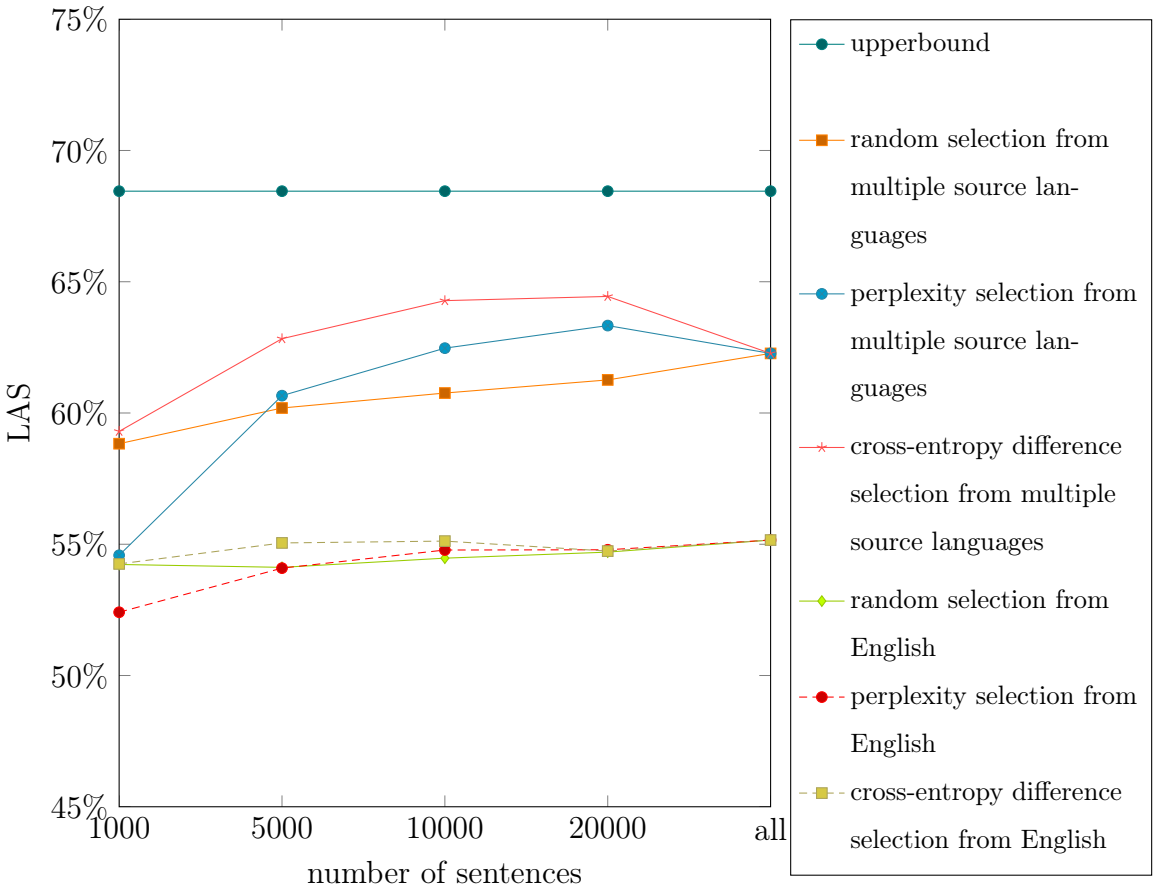

Figure 5.6: Labeled attachment scores (LAS) for Spanish with different instance selection methods

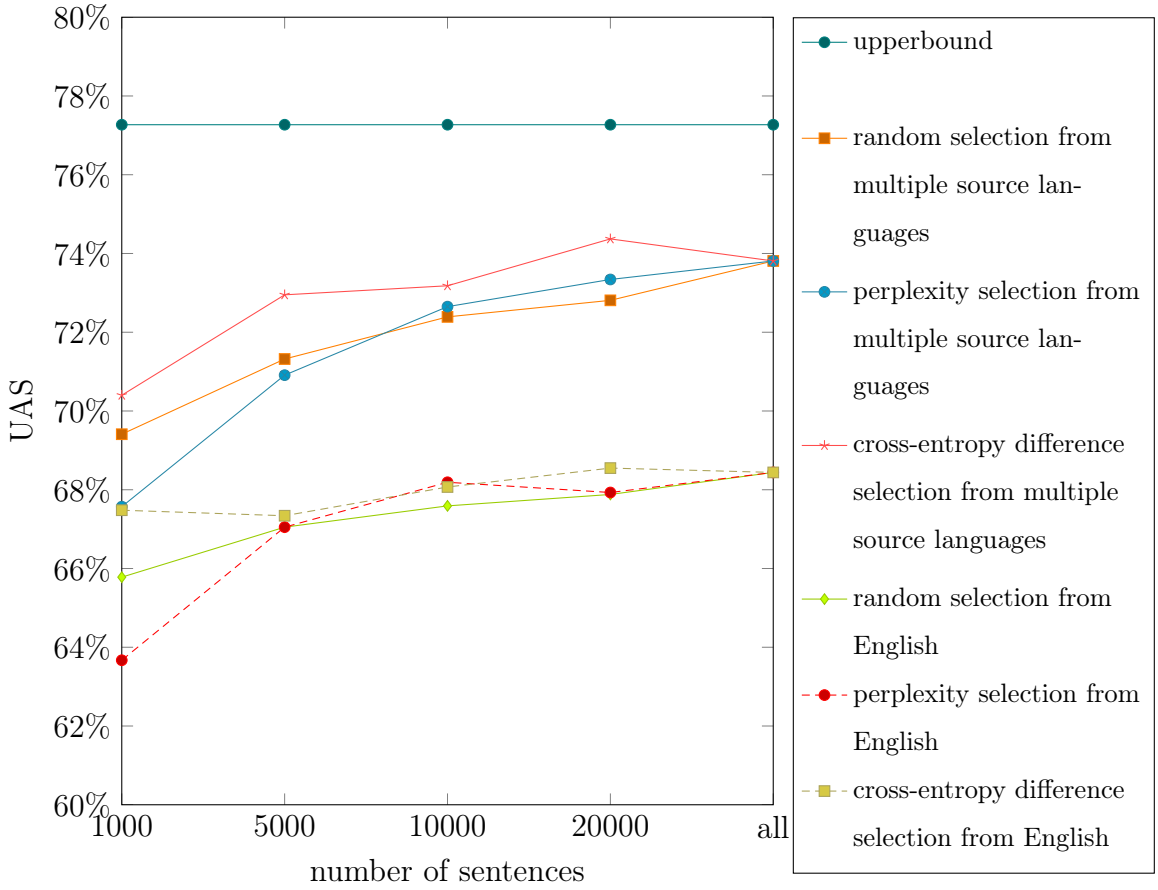

Figure 5.7: Unlabeled attachment scores (UAS) for French with different instance selection methods

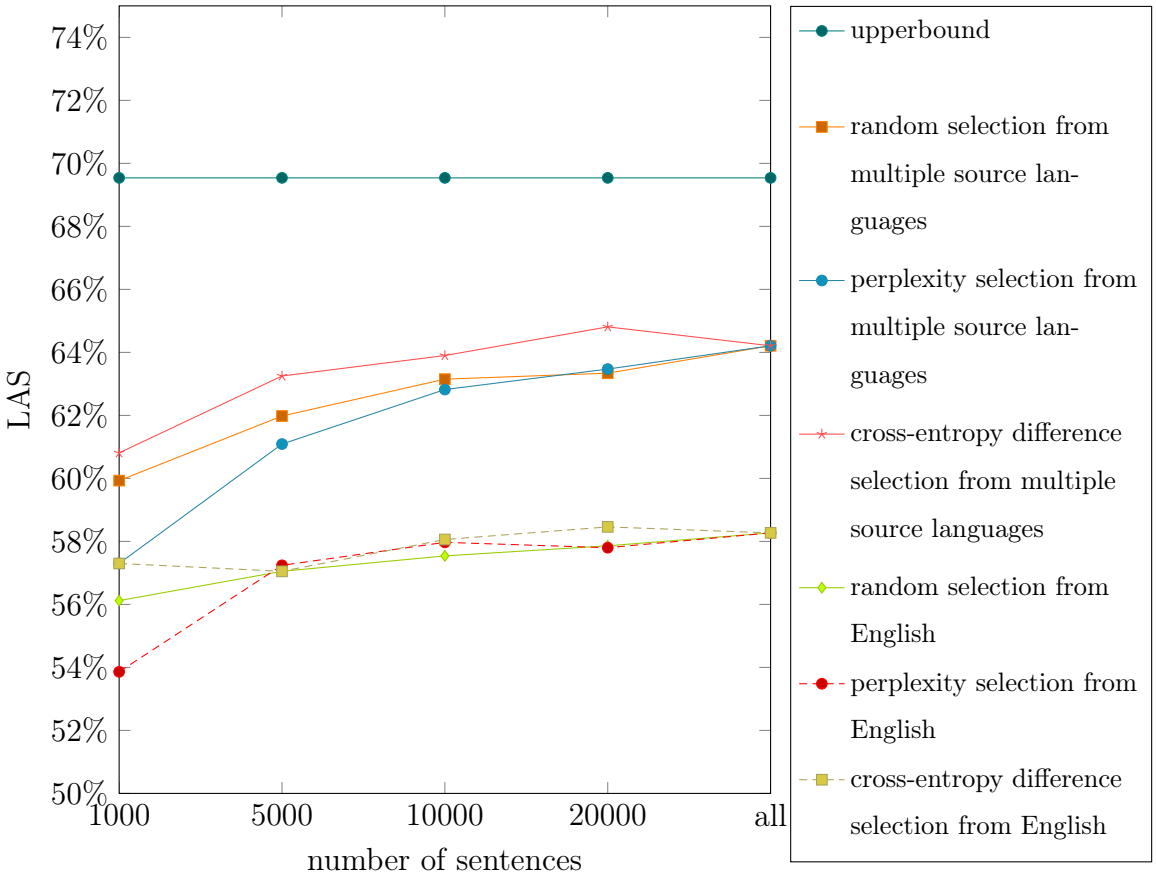

Figure 5.8: Labeled attachment scores (LAS) for French with different instance selection methods

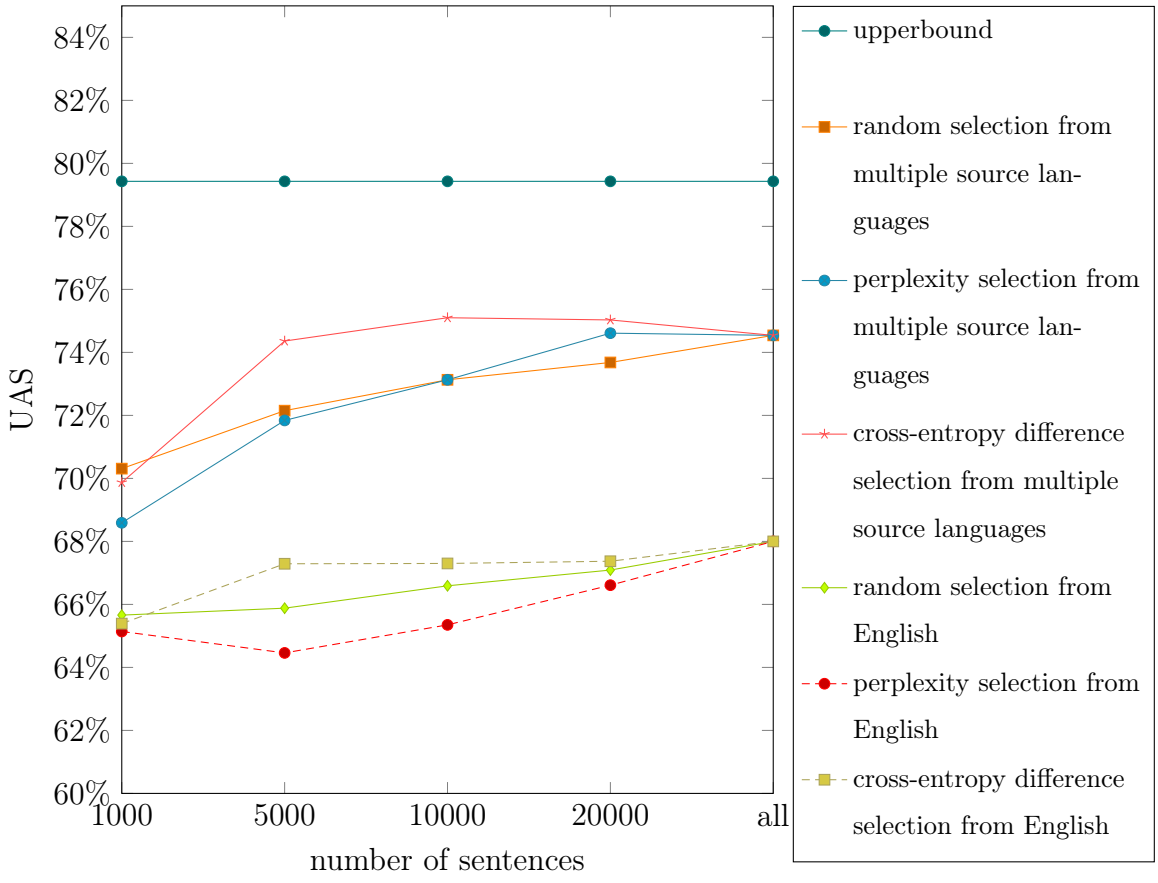

Figure 5.9: Unlabeled attachment scores (UAS) for Italian with different instance selection methods

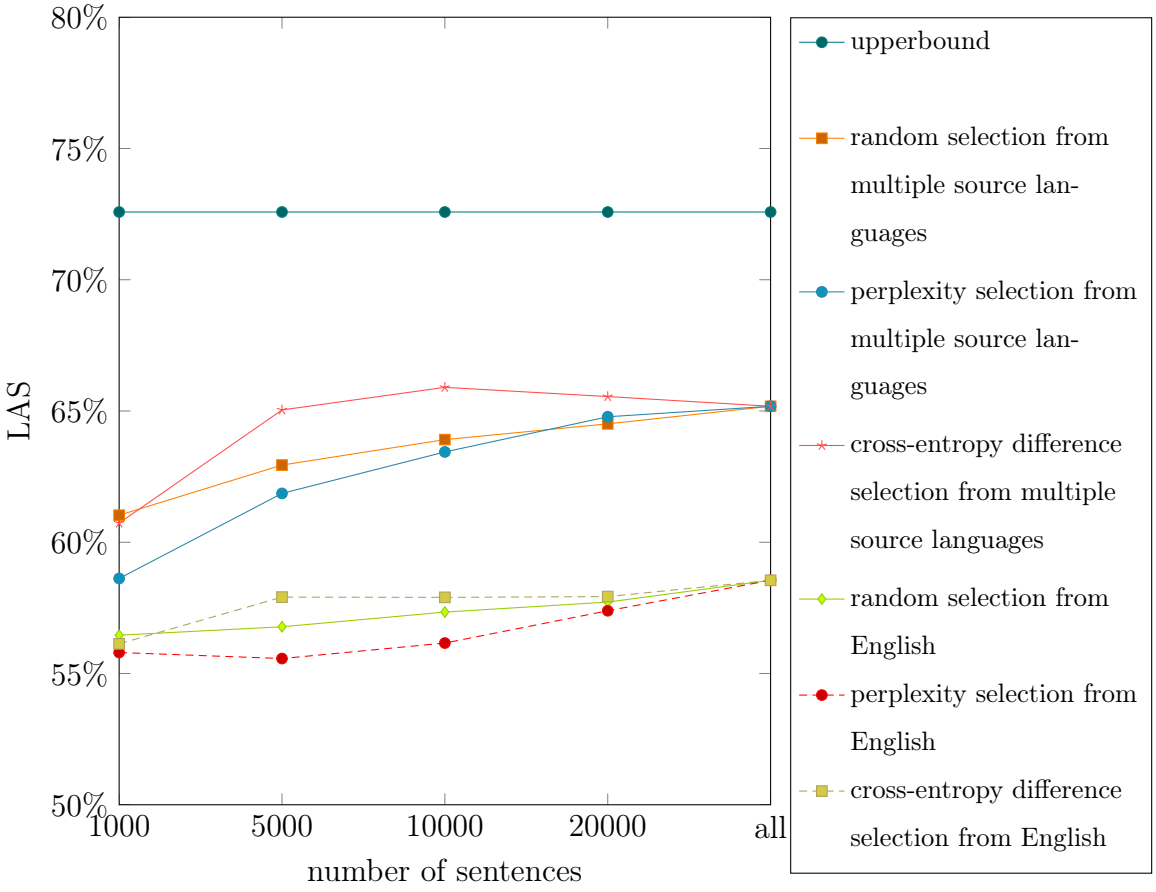

Figure 5.10: Labeled attachment scores (LAS) for Italian with different instance selection methods

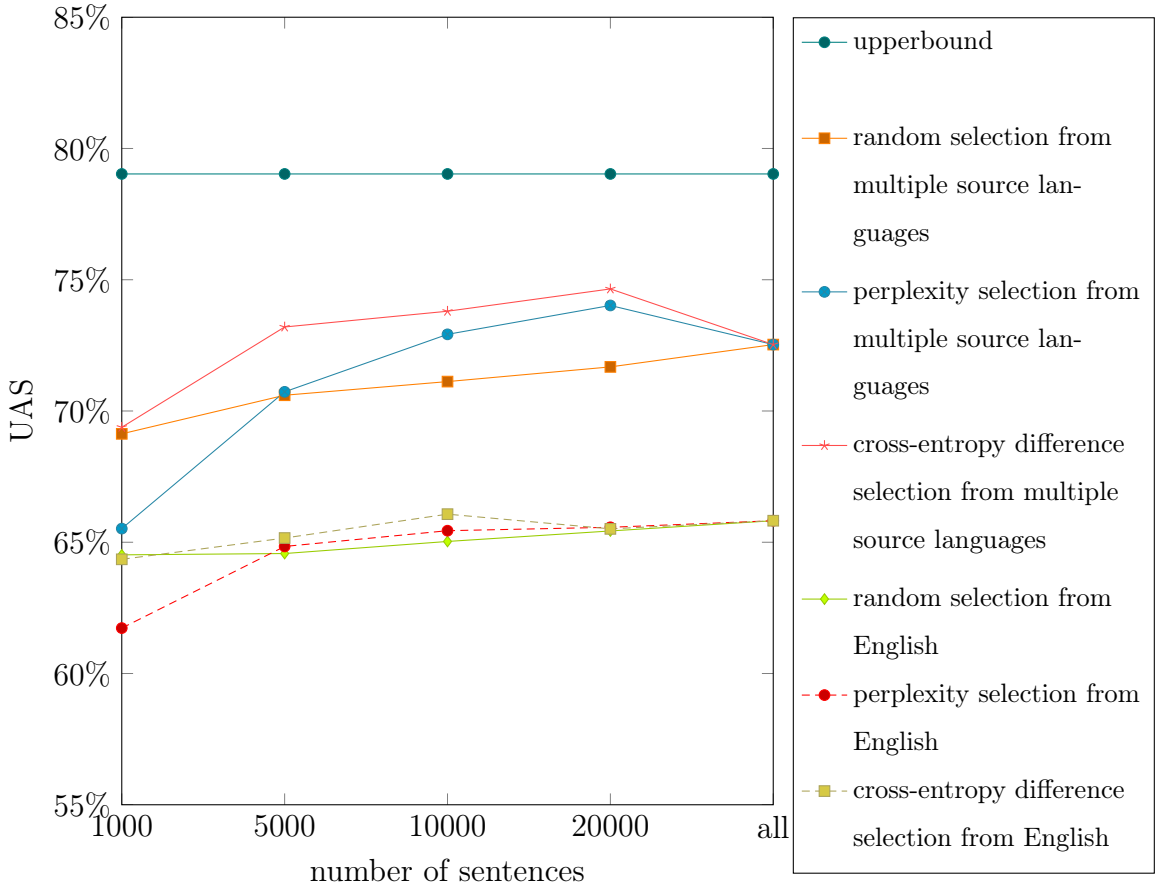

Figure 5.11: Unlabeled attachment scores (UAS) for Brazilian Portuguese with different instance selection methods

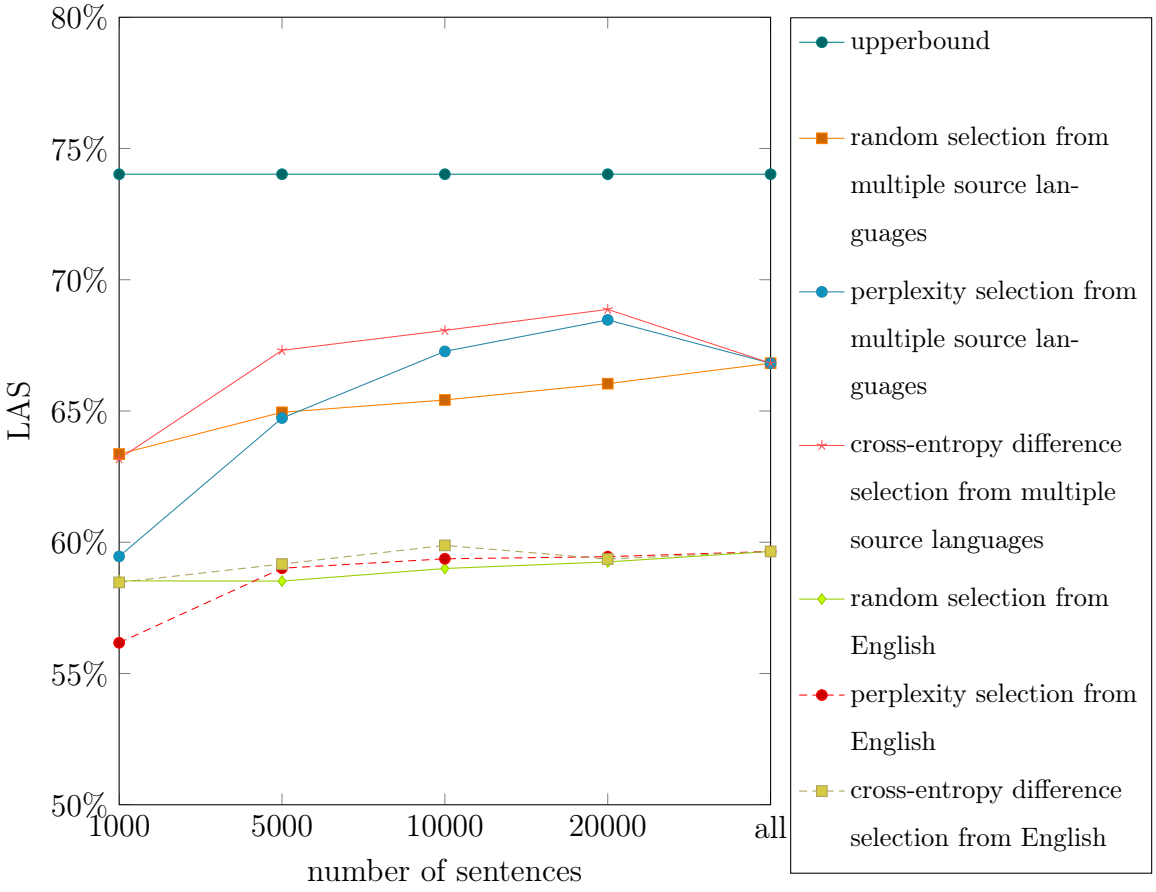

Figure 5.12: Labeled attachment scores (LAS) for Brazilian Portuguese with different instance selection methods

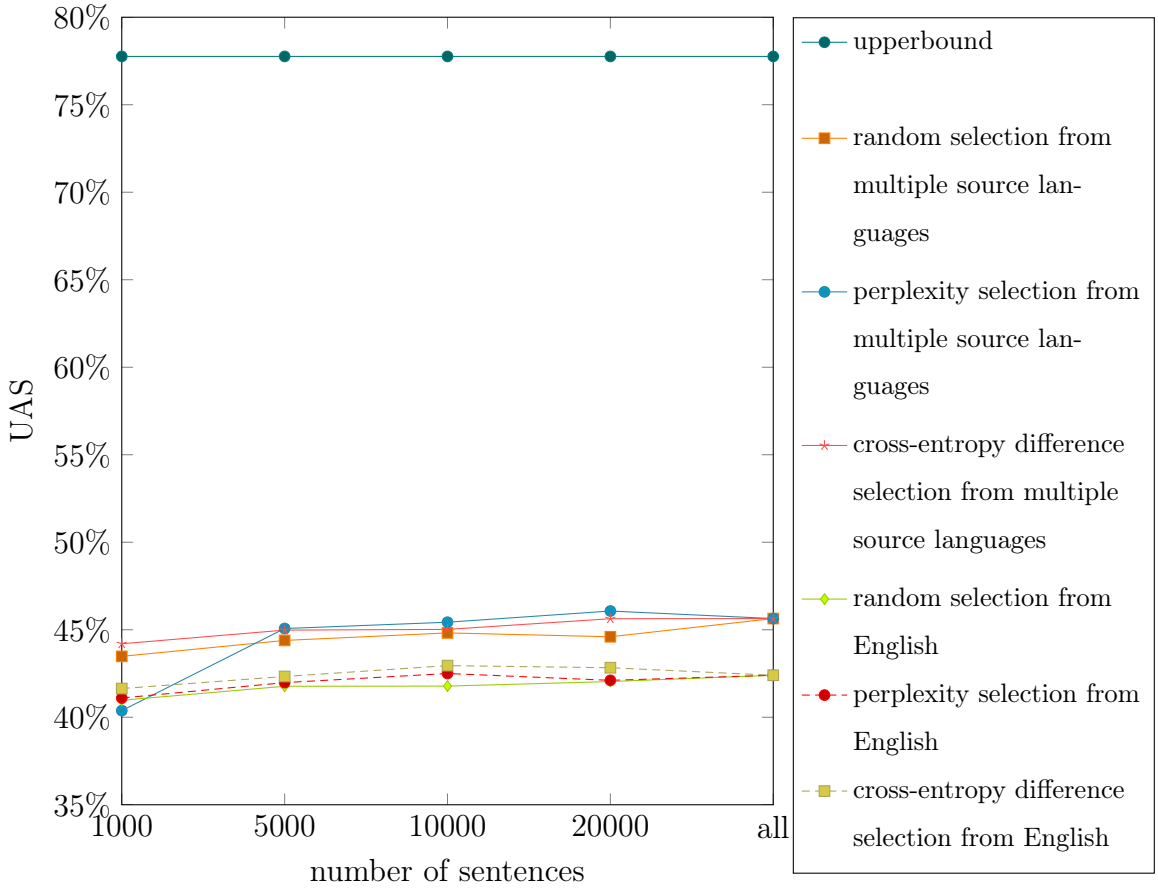

Figure 5.13: Unlabeled attachment scores (UAS) for Indonesian with different instance selection methods

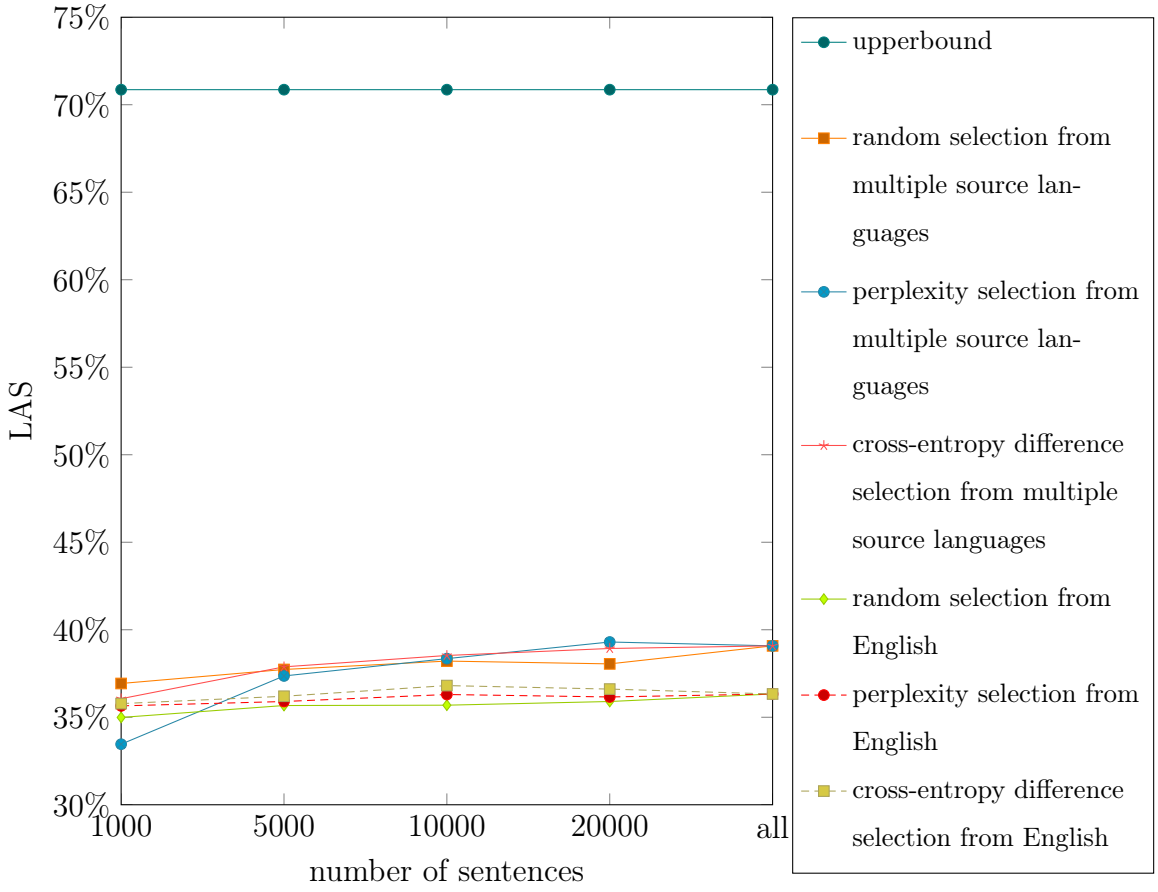

Figure 5.14: Labeled attachment scores (LAS) for Indonesian with different instance selection methods

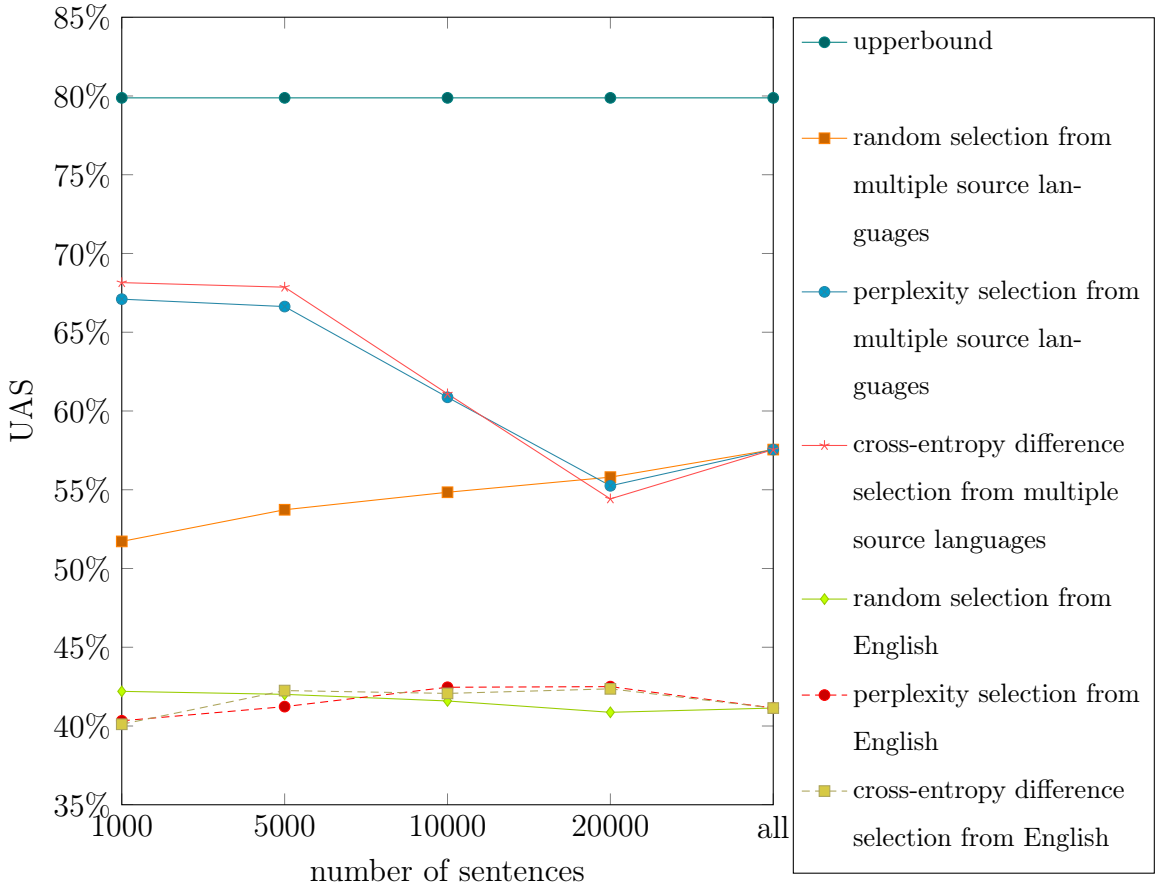

Figure 5.15: Unlabeled attachment scores (UAS) for Japanese with different instance selection methods

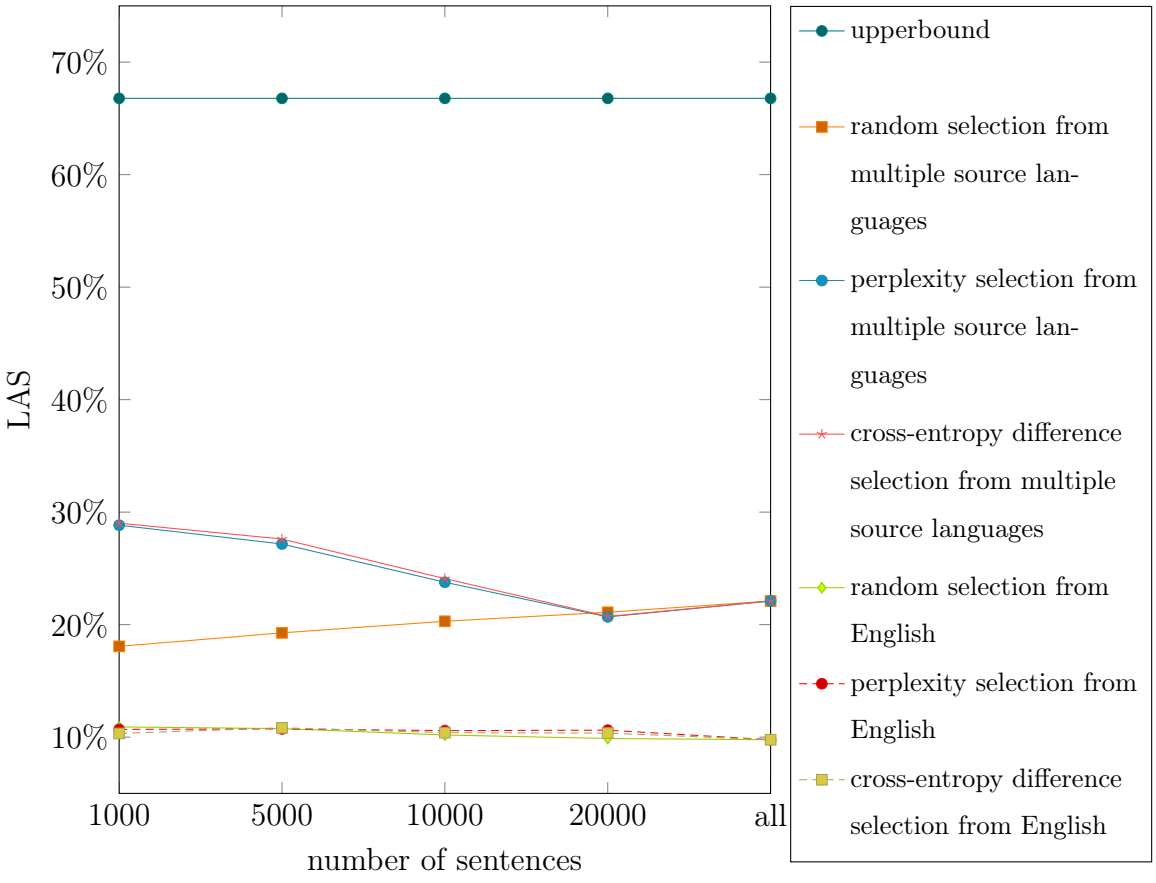

Figure 5.16: Labeled attachment scores (LAS) for Japanese with different instance selection methods

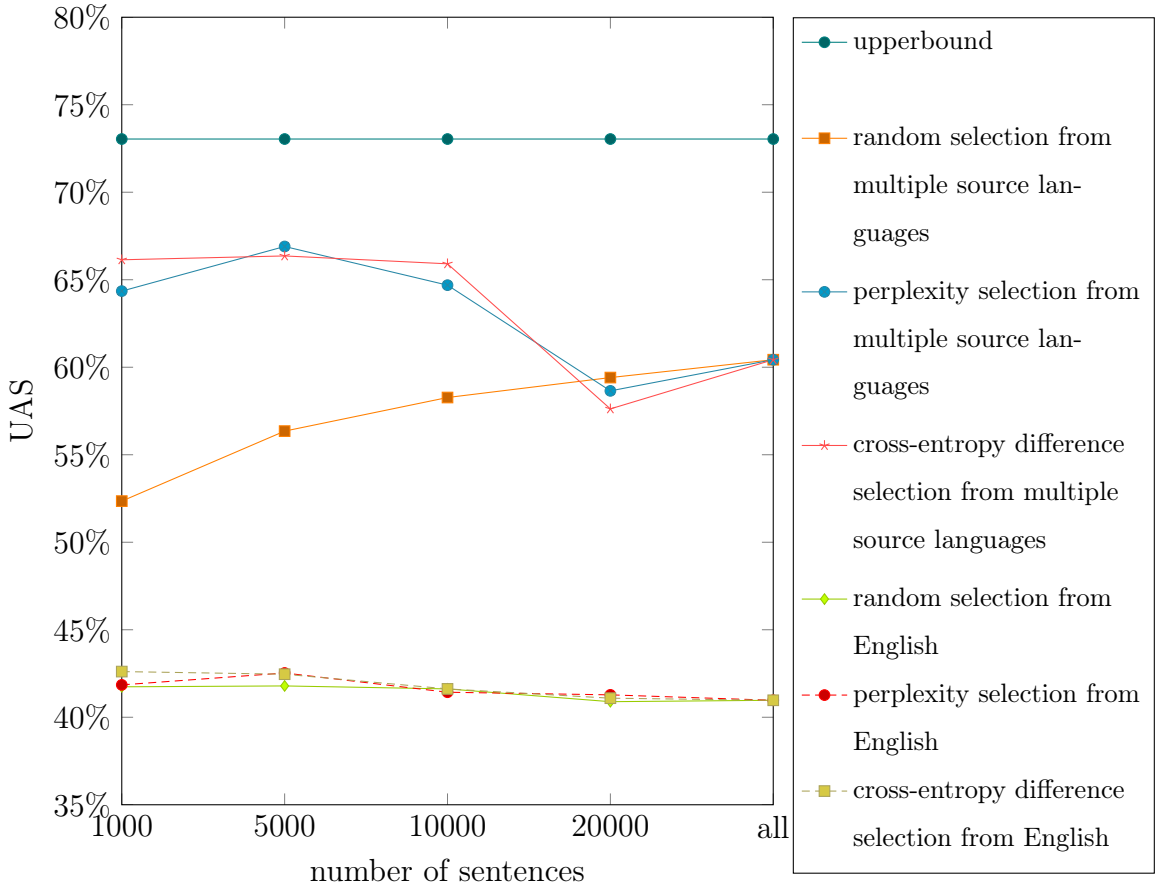

Figure 5.17: Unlabeled attachment scores (UAS) for Korean with different instance selection methods

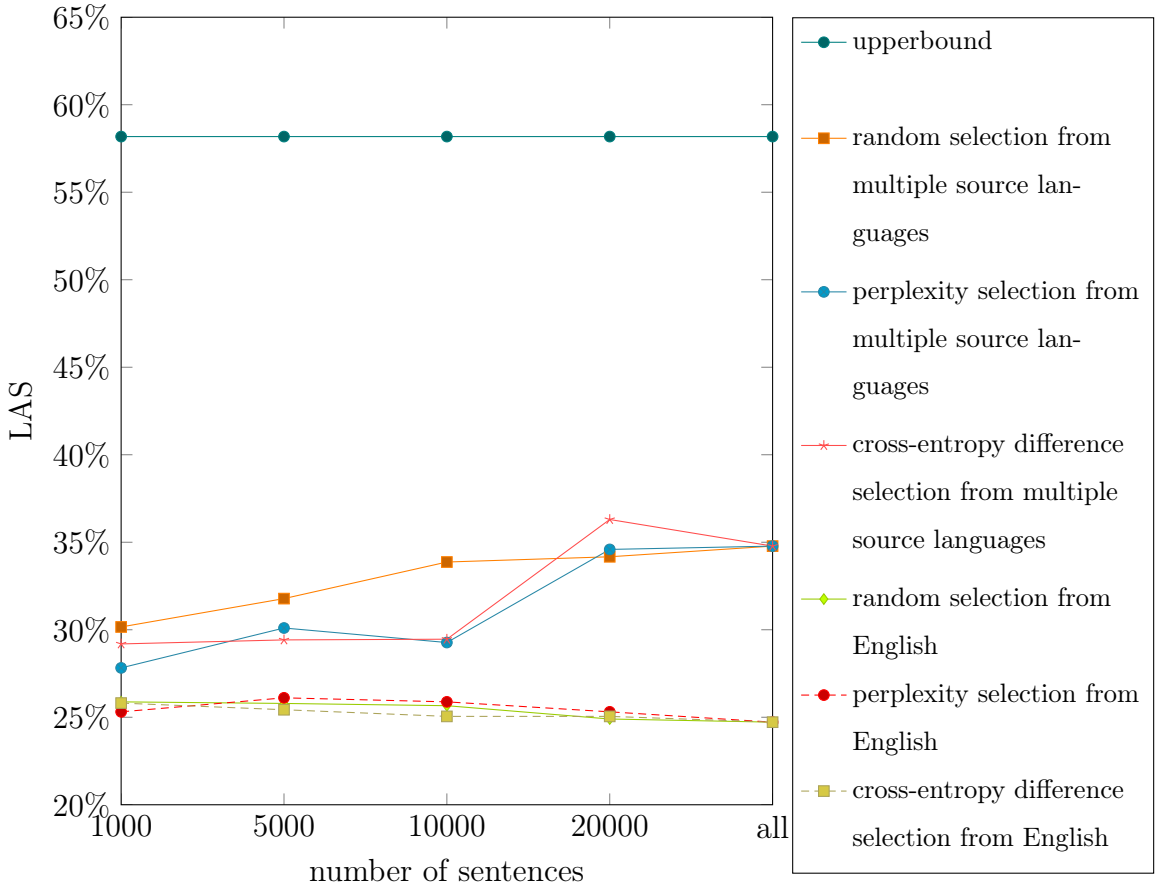

Figure 5.18: Labeled attachment scores (LAS) for Korean with different instance selection methods

perhaps not surprisingly, the number of sentences selected at each perplexity threshold varies greatly from language to language, as is shown in Table 5.3. At a cut-off of 6, between 20,000 and 30,000 training instances are selected for the Germanic languages, while more than 30,000 are selected for the Romance languages (with more than 40,000 for Italian). In stark contrast, only around 4,000 training instances are selected at this point for Indonesian and Korean and only around 2,000 for Japanese. This makes sense as the high quantity of training data for the four Romance languages makes it likely that there is a lot of highly relevant training instances to be selected for each of them from the remaining three, similarly with the three Germanic languages. By contrast, there is likely not much relevant training data available for Indonesian, Japanese, or Korean from the large Germanic and Romance languages' training datasets.

Among the ten languages tested, Japanese and Korean benefited most from this method of instance selection; this is likely due to each of these languages having relevant training data in the other one, as shown in the baseline numbers, amongst lots of irrelevant training data from the other source languages. However, the boost for these two languages, while quite large over using all of the training data, still provided scores lower than using the full training dataset of the other language (that is, the highest score for using perplexity selection to select relevant sentences from multiple source languages for Japanese as a target language is still lower than simply using only Korean training data). Tables 5.4 and 5.5 show the distribution of the source languages selected for Japanese and Korean. Overwhelmingly, perplexity selection for Korean selects mostly Japanese source sentences, while perplexity selection for Japanese selects mostly Korean source sentences.

This indicates that, although this method seems promising, if one has knowledge of the low-resource target language and can identify one (or more) similar source language(s), it is better to simply use direct transfer and train on the similar source language(s). However, this could still be useful for low-resource target languages

where not much is known or easily obtained about the grammar of the language, or to create a versatile pipeline that can be applied to multiple new languages as needed, without manual intervention or curation, although a POS tagged dataset needs to be available in the target language.

While perplexity selection is effective when using a pool of training data from multiple source languages, its performance when using only English training data is much more disappointing. Although occasional higher scores are seen, as in Japanese and Korean when using a perplexity threshold of 4 (for UAS), there are no consistent trends and no clear threshold that maximizes scores across languages, as when using multiple source languages. This may be because the English training data simply doesn't have relevant training data for some of the languages tested, or that the determination of relevance must be done using different methods. Perhaps this is unsurprising given that certain grammatical structures which are common in other languages will never, or very rarely, be seen in English. For instance, when abstracting to POS tags, a language with SOV word order and optional subjects is likely to have a noun verb POS tag sequence that is correctly interpreted as an object depending on a verb, but this same POS tag sequence in English is much more likely to be a subject depending on a verb, to the point where a dependency parser trained on English training data, regardless of selection method, is likely to never parse a noun verb POS tag sequence as being an object depending on a verb.

This disappointing performance may be due to a lack of coverage of more complex grammatical structures in these languages, since selecting by perplexity favors shorter, simpler sentences and a lower proportion of unique sentences (as the same sentence will always have the same perplexity against a given language model and if that perplexity is low, then all of the identical sentences will be included with a given selection).

Perplexity Cut-Off	Target Test Language									
	Number of Sentences Selected									
	Germanic			Romance						
	DE	**EN**	**SV**	**ES**	**FR**	**IT**	**PT-BR**	**ID**	**JA**	**KO**
3	39	32	90	313	339	244	299	131	330	637
4	2060	2222	2211	5251	4925	5626	5602	853	1145	2360
5	9723	10279	11186	18081	16805	22103	19479	4452	2125	4075
6	23244	23058	28213	33154	31200	42998	35762	13179	2922	5360
7	38505	36331	48008	46841	44026	61357	49715	26983	3642	6567
8	52405	47034	64903	58323	54904	74964	61070	42676	4370	7496
9	63460	54816	77258	67456	63780	83873	70217	57406	4916	8401
none	107108	81394	116779	107088	106715	114837	111626	116749	112949	115789

Table 5.3: The number of training sentences selected from a pool of multiple source languages with varying perplexity thresholds

Number of Sentences	Source Language								
	Number of Sentences Selected								
	Germanic			Romance					
	DE	**EN**	**SV**	**ES**	**FR**	**IT**	**PT-BR**	**ID**	**KO**
1,000	2	99	16	4	2	0	2	6	869
5,000	51	294	163	21	22	6	14	111	4318
10,000	749	2021	298	167	172	65	141	1088	5299
20,000	2684	6860	632	610	898	273	563	2082	5398
all	14118	39832	4447	14138	14511	6389	9600	4477	5437

Table 5.4: The number of sentences selected from each source language's training data when using perplexity selection for Japanese

Number of Sentences	Source Language								
	Number of Sentences Selected								
	Germanic			Romance					
	DE	**EN**	**SV**	**ES**	**FR**	**IT**	**PT-BR**	**ID**	**JA**
1,000	0	4	1	0	0	0	0	0	995
5,000	20	93	18	11	7	3	6	46	4796
10,000	305	1482	109	92	66	20	116	578	7232
20,000	1646	6884	543	494	428	184	593	1445	7783
all	14118	39832	4447	14138	14511	6389	9600	4477	8277

Table 5.5: The number of sentences selected from each source language's training data when using perplexity selection for Korean

5.2 Cross-Entropy Difference Selection

The results for cross-entropy difference selection are very similar to those seen with perplexity selection. Once again, this method seems somewhat effective when using multiple source languages, and a threshold can be used to come close to maximizing scores across languages (in this case, a cut-off of around 0). Additionally, the same methodology applied to only English training data fails to produce promising results.

5.3 Rearranging POS Tags

Rearranging the order of modifying adjectives relative to the nouns they modify in English training data to better match the word order seen in the target languages had a small positive impact on the results, which is promising for more extensive rearrangement. Instance selection methods applied to this rearranged data did not perform any better than on the non-rearranged data. Modifying the English training data in this way results in a slight drop in scores for Spanish over non-rearranged English, but there is a boost over non-rearranged English for French and Indonesian. For French, this boost is about half a percentage point for both UAS and LAS, while for Indonesian, it is about a full percentage point. It seems reasonable that Indonesian would have the biggest boost from this technique, since it overwhelmingly has the highest relative frequency of modifying adjectives following the nouns they modify. However, these scores are all lower than those using multiple source languages for training data.

Chapter 6

CONCLUSION

Dependency parsing is an important NLP task with many downstream applications, and as is common in the field, high accuracy results can be obtained when using statistical methods and training on high-quality annotated training data. When dealing with low-resource languages where annotated training data is not available and prohibitively expensive to obtain, more clever methods must be used to leverage existing resources. Most current methods leverage parallel data and projected transfer techniques to improve dependency parsing. My work in this thesis focuses instead on instance selection, which rests on the assumption, little explored cross-linguistically but well-proven monolingually in domain adaptation, that using less training data that is more relevant to your test case is better than using a full pool of potentially highly irrelevant training data. A benefit of this approach is eliminating the need for parallel data which is likely to be quite difficult to obtain for a true low-resource language, resting instead on part-of-speech (POS) tagged data in the target language.

In this thesis, I presented results for two different methods of instance selection, which showed varied results. Both methods rely on a language model trained on POS tags for the target language. Perplexity selection uses this language model to calculate perplexity for each training data instance and select those with the lowest perplexity. Cross-entropy difference selection uses an additional language model, trained on the source training data, and the cross-entropy of each training data instance is calculated against both language models, and the difference of the two cross-entropies are taken and used to select those with the lowest cross-entropy difference. These methods were applied to three different pools of training data, the first being a pool of multiple

source languages, the second being English data only, and the third being a rearranged version of English where all modifying adjectives preceding nouns were rearranged to follow the nouns they modify, in order to provide a closer match to languages with different word order.

When a pool of multiple source languages is used, a significant boost to both unlabeled attachment scores (UAS) and labeled attachment scores (LAS) are seen for target languages where relevant training data is available but infrequent in the training data, with cross-entropy difference providing very slightly better performance over perplexity selection. These selection methods don't provide the same large improvements for target languages with lots of relevant training data available among the multiple source languages, but they also have no negative impact. When only English is used for the training data, the results are much less clear; while some boost is seen for some languages, the impact is minimal. Rearranging the order of modifying adjectives relative to the nouns they modify in English training data to better match the word order seen in the target languages had a small positive impact on the results, which is promising for more extensive rearrangement. Instance selection methods applied to this rearranged data did not perform any better than on the non-rearranged data.

While the methods tested here did not provide state of the art performance for dependency parsing for low-resource languages, they may pave the way for future exploration in this space.

Chapter 7
FUTURE WORK

There are many clear options for future work building on the results of this research. Further exploration of instance selection, automated rearrangement of POS tags, and the downstream impact of these methods could all prove fruitful.

Although these methods of instance selection had mixed success, other methods of instance selection may yield greater success. Leveraging POS tags in a different way or using something other than POS tags to identify relevant training instances could be a possible path of study. Additionally, changing the selection process to select for coverage, rather than simply selecting the most similar instances, could help to account for as many different phrases and ordering of phrases as possible.

Despite a lack of training data, low-resource languages often have documented information on word order and phrase structures that could be leveraged. Further exploration into rearranging source language(s) may provide a greater boost than seen here, especially if long distance rearrangement is explored. Automatically detecting rearrangement candidates based on POS tag sequence frequencies could also be useful.

Lastly, testing the same methodology explored here in a pipeline could bring informative results and help with further error analysis. Gold POS tags were used for the purposes of this research, but in a real world situation, these may not be accessible, and automatic POS tags would likely be used instead. Exploring the impact automatic POS tagging has on the dependency parse on the resulting output would be very interesting. Additionally, examining the impact of the different dependency parsers on a downstream application, such as semantic analysis, machine translation, or question answering, could be an informative method of extrinsic evaluation.

BIBLIOGRAPHY

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics, 2011.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. Converting Italian treebanks: Towards an Italian Stanford Dependency Treebank. *LAW VII & ID*, page 61, 2013.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 449–454, 2006.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, volume 6, pages 449–454, 2006.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. Dependency grammar induction via bitext projection constraints. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 369–377, Suntec, Singapore, August 2009. ACL and AFNLP.

Ryan Georgi, Fei Xia, and William D. Lewis. Enhanced and portable dependency projection algorithms using interlinear glossed text. In *Proceedings of the 51st*

Annual Meeting of the Association for Computational Linguistics, pages 306–311, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. Building the essential resources for Finnish: the Turku Dependency Treebank. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1–39. Springer, 2013.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 1:1–15, 2004.

Ryan McDonald, Koby Crammer, and Fernando Pereira. Online large-margin training of dependency parsers. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 91–98. Association for Computational Linguistics, 2005a.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan HajiÄIↃ. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–530, 2005b.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 216–220. Association for Computational Linguistics, 2006.

Ryan McDonald, Slav Petrov, and Keith Hall. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods*

in Natural Language Processing (EMNLP), pages 62–72, Edinburgh, Scotland, UK, July 2011. Association for Computational Linguistics.

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

Ryan T McDonald and Fernando CN Pereira. Online learning of approximate dependency parsing algorithms. In *11th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006*, pages 81–88, 2006.

Robert C Moore and William Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224. Association for Computational Linguistics, 2010.

Joakim Nivre, Joham Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Sesson of EMNLP-CoNLL 2007*, pages 915–932, Prague, June 2007. Association for Computational Linguistics.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2012.

David A. Smith and Jason Eisner. Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 822–831, Singapore, August 2009. ACL and AFNLP.

Anders Søgaard. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics:shortpapers*, pages 682–686, Portland, Oregon, June 2011. Association for Computational Linguistics.

Anders Søgaard and Martin Haulrich. Sentence-level instance-weighting for graph-based and transition-based dependency parsing. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 43–47, Dublin City University, October 2011. Association for Computational Linguistics.

Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto, and Eiichiro Sumita. Method of selecting training data to build a compact and efficient translation model. In *IJCNLP*, pages 655–660, 2008.

Daniel Zeman and Philip Resnik. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42, Hyderabad, India, January 2008. Asian Federation of Natural Language Processing.