Effective Measures of Domain Similarity for Parsing

Barbara Plank
University of Groningen
The Netherlands
b.plank@rug.nl

Gertjan van Noord
University of Groningen
The Netherlands
G.J.M.van.Noord@rug.nl

Abstract

It is well known that parsing accuracy suffers when a model is applied to out-of-domain data. It is also known that the most beneficial data to parse a given domain is data that matches the domain (Sekine, 1997; Gildea, 2001). Hence, an important task is to select appropriate domains. However, most previous work on domain adaptation relied on the implicit assumption that domains are somehow given. As more and more data is becoming available, automatic ways to select data that is beneficial for a new (unknown) target domain are becoming attractive. This paper evaluates various ways to automatically acquire related training data for a given test set. The results show that an unsupervised technique based on topic models is effective – it outperforms random data selection on both examined languages, English and Dutch. Moreover, the technique works better than manually assigned labels gathered from meta-data that is available for English.

1 Introduction and Motivation

Previous research on domain adaptation has focused on the task of adapting a system trained on one domain, say newspaper text, to a particular new domain, say biomedical data. Usually, some amount of (labeled or unlabeled) data from the new domain was given – which has been determined by a human.

However, with the growth of the web, more and more data is becoming available, where each document "is potentially its own domain" (McClosky et al., 2010). It is not straightforward to determine which data or model (in case we had several source domain models) would perform best on a new (unknown) target domain. Therefore, an important issue that arises is *how to measure domain similarity*, i.e. whether we can find a simple yet effective method to determine which model or data is most beneficial for an arbitrary piece of new text. Moreover, if we had such a measure, a related question is whether it can tell us something more about what is actually meant by "domain". So far, it was mostly arbitrarily used to refer to some kind of coherent unit (related to topic, style or genre), e.g.: newspaper text, biomedical abstracts, questions, fiction.

Most previous work on domain adaptation, for instance Hara et al. (2005), McClosky et al. (2006), Blitzer et al. (2006), Daumé III (2007), sidestepped this problem of automatic domain selection and adaptation. For parsing, to our knowledge only one recent study has started to examine this issue (McClosky et al., 2010) – we will discuss their approach in Section 2. Rather, an implicit assumption of all of these studies is that domains are given, i.e. that they are represented by the respective corpora. Thus, a corpus has been considered a homogeneous unit. As more data is becoming available, it is unlikely that domains will be 'given'. Moreover, a given corpus might not always be as homogeneous as originally thought (Webber, 2009; Lippincott et al., 2010). For instance, recent work has shown that the well-known Penn Treebank (PT) Wall Street Journal (WSJ) actually contains a variety of genres, including letters, wit and short verse (Webber, 2009).

In this study we take a different approach. Rather than viewing a given corpus as a monolithic entity,

we break it down into the article-level, and disregard corpora boundaries. Given the resulting set of documents (articles), we evaluate various ways to automatically acquire related training data for a given test set, to find answers to the following questions:

- Given a pool of data (a collection of articles from unknown domain) and a test article, is there a way to automatically select data that is relevant for the new domain? If so:
- Which similarity measure is good for parsing?
- How does it compare to human-annotated data?
- Is the measure also useful for other languages and/or tasks?

To this end, we evaluate measures of domain similarity and feature representations and their impact on dependency parsing accuracy. Given a collection of annotated articles, and a new article that we want to parse, we want to select the most similar articles to train the best parser for that new article.

In the following, we will first compare automatic measures to human-annotated labels by examining parsing performance within subdomains of the Penn Treebank WSJ. Then, we extend the experiments to the domain adaptation scenario. Experiments were performed on two languages: English and Dutch. The empirical results show that a simple measure based on topic distributions is effective for both languages and works well also for Part-of-Speech tagging. As the approach is based on plain surface-level information (words) and it finds related data in a completely unsupervised fashion, it can be easily applied to other tasks or languages for which annotated (or automatically annotated) data is available.

2 Related Work

The work most related to ours is McClosky et al. (2010). They try to find the best combination of source models to parse data from a new domain, which is related to Plank and Sima'an (2008). In the latter, unlabeled data was used to create several parsers by weighting trees in the WSJ according to their similarity to the subdomain. McClosky et al. (2010) coined the term *multiple source domain adaptation*. Inspired by work on parsing accuracy prediction (Ravi et al., 2008), they train a linear regression model to predict the best (linear interpolation) of source domain models. Similar to us, McClosky et al. (2010) regard a target domain as mixture of source domains, but they focus on phrase-structure parsing. Furthermore, our approach differs to theirs in two respects: we do not treat source corpora as one entity and try to mix models, but rather consider articles as base units and try to find subsets of related articles (the most similar articles); moreover, instead of creating a supervised model (in their case to predict parsing accuracy), our approach is 'simplistic': we apply measures of domain similarity directly (in an unsupervised fashion), without the necessity to train a supervised model.

Two other related studies are (Lippincott et al., 2010; Van Asch and Daelemans, 2010). Van Asch and Daelemans (2010) explore a measure of domain difference (Renyi divergence) between pairs of domains and its correlation to Part-of-Speech tagging accuracy. Their empirical results show a linear correlation between the measure and the performance loss. Their goal is different, but related: rather than finding related data for a new domain, they want to estimate the loss in accuracy of a PoS tagger when applied to a new domain. We will briefly discuss results obtained with the Renyi divergence in Section 5.1. Lippincott et al. (2010) examine subdomain variation in biomedicine corpora and propose awareness of NLP tools to such variation. However, they did not yet evaluate the effect on a practical task, thus our study is somewhat complementary to theirs.

The issue of data selection has recently been examined for Language Modeling (Moore and Lewis, 2010). A subset of the available data is automatically selected as training data for a Language Model based on a scoring mechanism that compares cross-entropy scores. Their approach considerably outperformed random selection and two previous proposed approaches both based on perplexity scoring.[1]

3 Measures of Domain Similarity

3.1 Measuring Similarity Automatically

Feature Representations A similarity function may be defined over any set of events that are con-

[1] We tested data selection by perplexity scoring, but found the Language Models too small to be useful in our setting.

sidered to be relevant for the task at hand. For parsing, these might be words, characters, n-grams (of words or characters), Part-of-Speech (PoS) tags, bilexical dependencies, syntactic rules, etc. However, to obtain more abstract types such as PoS tags or dependency relations, one would first need to gather respective labels. The necessary tools for this are again trained on particular corpora, and will suffer from domain shifts, rendering labels noisy.

Therefore, we want to gauge the effect of the simplest representation possible: plain surface characteristics (unlabeled text). This has the advantage that we do not need to rely on additional supervised tools; moreover, it is interesting to know how far we can get with this level of information only.

We examine the following feature representations: relative frequencies of words, relative frequencies of character tetragrams, and topic models, motivated as follows. Relative frequencies of words are a simple and effective representation used e.g. in text classification (Manning and Schütze, 1999), while character n-grams have proven successful in genre classification (Wu et al., 2010). Topic models (Blei et al., 2003; Steyvers and Griffiths, 2007) can be considered an advanced model over word distributions: every article is represented by a topic distribution, which in turn is a distribution over words. Similarity between documents can be measured by comparing topic distributions.

Similarity Functions There are many possible similarity (or distance) functions. They fall broadly into two categories: probabilistically-motivated and geometrically-motivated functions. The similarity functions examined in this study will be described in the following.

The *Kullback-Leibler (KL) divergence* $D(q||r)$ is a classical measure of 'distance'[2] between two probability distributions, and is defined as: $D(q||r) = \sum_y q(y) \log \frac{q(y)}{r(y)}$. It is a non-negative, additive, asymmetric measure, and 0 iff the two distributions are identical. However, the KL-divergence is undefined if there exists an event y such that $q(y) > 0$ but $r(y) = 0$, which is a property that "makes it unsuitable for distributions derived via maximum-likelihood estimates" (Lee, 2001).

[2]It is not a proper distance metric since it is asymmetric.

One option to overcome this limitation is to apply smoothing techniques to gather non-zero estimates for all y. The alternative, examined in this paper, is to consider approximations to the KL divergence, such as the Jensen-Shannon (JS) divergence (Lin, 1991) and the skew divergence (Lee, 2001).

The *Jensen-Shannon divergence*, which is symmetric, computes the KL-divergence between q, r, and the average between the two. We use the JS divergence as defined in Lee (2001): $JS(q,r) = \frac{1}{2}[D(q||\text{avg}(q,r)) + D(r||\text{avg}(q,r))]$. The asymmetric *skew divergence* s_α, proposed by Lee (2001), mixes one distribution with the other by a degree defined by $\alpha \in [0,1)$: $s_\alpha(q,r,\alpha) = D(q||\alpha r + (1-\alpha)q)$. As α approaches 1, the skew divergence approximates the KL-divergence.

An alternative way to measure similarity is to consider the distributions as vectors and apply geometrically-motivated distance functions. This family of similarity functions includes the *cosine* $cos(q,r) = q(y) \cdot r(y)/||q(y)||||r(y)||$, *euclidean* $euc(q,r) = \sqrt{\sum_y (q(y) - r(y))^2}$ and *variational* (also known as L1 or Manhattan) distance function, defined as $var(q,r) = \sum_y |q(y) - r(y)|$.

3.2 Human-annotated data

In contrast to the automatic measures devised in the previous section, we might have access to human annotated data. That is, use label information such as topic or genre to define the set of similar articles.

Genre For the Penn Treebank (PT) Wall Street Journal (WSJ) section, more specifically, the subset available in the Penn Discourse Treebank, there exists a partition of the data by *genre* (Webber, 2009). Every article is assigned one of the following genre labels: news, letters, highlights, essays, errata, wit and short verse, quarterly progress reports, notable and quotable. This classification has been made on the basis of meta-data (Webber, 2009). It is well-known that there is no meta-data directly associated with the individual WSJ files in the Penn Treebank. However, meta-data can be obtained by looking at the articles in the ACL/DCI corpus (LDC99T42), and a mapping file that aligns document numbers of DCI (DOCNO) to WSJ keys (Webber, 2009). An example document is given in Figure 1. The meta-data field `HL` contains headlines, `SO` source info, and

the IN field includes topic markers.

```
<DOC><DOCNO>  891102-0186. </DOCNO>
<WSJKEY> wsj_0008 </WSJKEY>
<AN> 891102-0186. </AN>
<HL> U.S. Savings Bonds Sales
@  Suspended by Debt Limit </HL>
<DD> 11/02/89 </DD>
<SO> WALL STREET JOURNAL (J) </SO>
<IN> FINANCIAL, ACCOUNTING, LEASING (FIN)
BOND MARKET NEWS (BON) </IN>
<GV> TREASURY DEPARTMENT (TRE) </GV>
<DATELINE> WASHINGTON  </DATELINE>
<TXT>
<p><s>
The federal government suspended sales of U.S.
savings bonds because Congress hasn't lifted
the ceiling on government debt.</s></p> [...]
```

Figure 1: Example of ACL/DCI article. We have augmented it with the WSJ filename (WSJKEY).

Topic On the basis of the same meta-data, we devised a classification of the Penn Treebank WSJ by *topic*. That is, while the genre division has been mostly made on the basis of headlines, we use the information of the IN field. Every article is assigned one, more than one or none of a predefined set of keywords. While their origin remains unclear,[3] these keywords seem to come from a controlled vocabulary. There are 76 distinct topic markers. The three most frequent keywords are: TENDER OFFERS, MERGERS, ACQUISITIONS (TNM), EARNINGS (ERN), STOCK MARKET, OFFERINGS (STK). This reflects the fact that a lot of articles come from the financial domain. But the corpus also contains articles from more distant domains, like MARKETING, ADVERTISING (MKT), COMPUTERS AND INFORMATION TECHNOLOGY (CPR), HEALTH CARE PROVIDERS, MEDICINE, DENTISTRY (HEA), PETROLEUM (PET).

4 Experimental Setup

4.1 Tools & Evaluation

The parsing system used in this study is the MST parser (McDonald et al., 2005), a state-of-the-art data-driven graph-based dependency parser. It is a system that can be trained on a variety of languages given training data in CoNLL format (Buchholz and Marsi, 2006). Additionally, the parser implements both projective and non-projective parsing algorithms. The projective algorithm is used for the experiments on English, while the non-projective variant is used for Dutch. We train the parser using default settings. MST takes PoS-tagged data as input; we use gold-standard tags in the experiments.

We estimate topic models using Latent Dirichlet Allocation (Blei et al., 2003) implemented in the MALLET[4] toolkit. Like Lippincott et al. (2010), we set the number of topics to 100, and otherwise use standard settings (no further optimization). We experimented with the removal of stopwords, but found no deteriorating effect while keeping them. Thus, all experiments are carried out on data where stopwords were not removed.

We implemented the similarity measures presented in Section 3.1. For skew divergence, that requires parameter α, we set $\alpha = .99$ (close to KL divergence) since that has shown previously to work best (Lee, 2001). Additionally, we evaluate the approach on English PoS tagging using two different taggers: MXPOST, the MaxEnt tagger of Ratnaparkhi[5] and Citar,[6] a trigram HMM tagger.

In all experiments, parsing performance is measured as *Labeled Attachment Score* (LAS), the percentage of tokens with correct dependency edge and label. To compute LAS, we use the CoNLL 2007 evaluation script[7] with punctuation tokens excluded from scoring (as was the default setting in CoNLL 2006). PoS tagging accuracy is measured as the percentage of correctly labeled words out of all words. Statistical significance is determined by *Approximate Randomization Test* (Noreen, 1989; Yeh, 2000) with 10,000 iterations.

4.2 Data

English - WSJ For English, we use the portion of the Penn Treebank Wall Street Journal (WSJ) that has been made available in the CoNLL 2008 shared

[3] It is not known what IN stands for, as also stated in Mark Liberman's notes in the readme of the ACL/DCI corpus. However, a reviewer suggested that IN might stand for "index terms" which seems plausible.

[4] http://mallet.cs.umass.edu/

[5] ftp://ftp.cis.upenn.edu/pub/adwait/jmx/

[6] Citar has been implemented by Daniël de Kok and is available at: https://github.com/danieldk/citar

[7] http://nextens.uvt.nl/depparse-wiki/

task. This data has been automatically converted[8] into dependency structure, and contains three files: the training set (sections 02-21), development set (section 24) and test set (section 23).

Since we use articles as basic units, we actually split the data to get back original article boundaries.[9] This led to a total of 2,034 articles (1 million words). Further statistics on the datasets are given in Table 1. In the first set of experiments on WSJ subdomains, we consider articles from section 23 and 24 that contain at least 50 sentences as test sets (target domains). This amounted to 22 test articles.

	EN: WSJ	WSJ+G+B	Dutch
articles	2,034	3,776	51,454
sentences	43,117	77,422	1,663,032
words	1,051,997	1,784,543	20,953,850

Table 1: Overview of the datasets for English and Dutch.

To test whether we have a reasonable system, we performed a sanity check and trained the MST parser on the training section (02-21). The result on the standard test set (section 23) is identical to previously reported results (excluding punctuation tokens: LAS 87.50, Unlabeled Attachment Score (UAS) 90.75; with punctuation tokens: LAS 87.07, UAS 89.95). The latter has been reported in (Surdeanu and Manning, 2010).

English - Genia (G) & Brown (B) For the Domain Adaptation experiments, we added 1,552 articles from the GENIA[10] treebank (biomedical abstracts from Medline) and 190 files from the Brown corpus to the pool of data. We converted the data to CoNLL format with the LTH converter (Johansson and Nugues, 2007). The size of the test files is, respectively: Genia 1,360 sentences with an average number of 26.20 words per sentence; the Brown test set is the same as used in the CoNLL 2008 shared task and contains 426 sentences with a mean of 16.80 words.

[8]Using the LTH converter: `http://nlp.cs.lth.se/software/treebank_converter/`

[9]This was a non-trivial task, as we actually noticed that some sentences have been omitted from the CoNLL 2008 shared task.

[10]We use the GENIA distribution in Penn Treebank format available at `http://bllip.cs.brown.edu/download/genia1.0-division-rel1.tar.gz`

5 Experiments on English

5.1 Experiments within the WSJ

In the first set of experiments, we focus on the WSJ and evaluate the similarity functions to gather related data for a given test article. We have 22 WSJ articles as test set, sampled from sections 23 and 24. Regarding feature representations, we examined three possibilities: relative frequencies of words, relative frequencies of character tetragrams (both unsmoothed) and document topic distributions.

In the following, we only discuss representations based on words or topic models as we found character tetragrams less stable; they performed sometimes similar to their word-based counterparts but other times, considerably worse.

Results of Similarity Measures Table 2 compares the effect of the different ways to select related data in comparison to the random baseline for increasing amounts of training data. The table gives the average over 22 test articles (rather than showing individual tables for the 22 articles). We select articles up to various thresholds that specify the total number of sentences selected in each round (e.g. 0.3k, 1.2k, etc.).[11] In more detail, Table 2 shows the result of applying various similarity functions (introduced in Section 3.1) over the two different feature representations (w: words; tm: topic model) for increasing amounts of data. We additionally provide results of using the Renyi divergence.[12]

Clearly, as more and more data is selected, the differences become smaller, because we are close to the data limit. However, for all data points less than 38k (97%), selection by jensen-shannon, variational and cosine similarity outperform random data selection significantly for both types of feature representations (words and topic model). For selection by topic models, this additionally holds for the euclidean measure.

From the various measures we can see that selection by jensen-shannon divergence and variational distance perform best, followed by cosine similarity, skew divergence, euclidean and renyi.

[11]Rather than choosing k articles, as article length may differ.

[12]The *Renyi divergence* (Rényi, 1961), also used by Van Asch and Daelemans (2010), is defined as $D_\alpha(q,r) = 1/(\alpha - 1)\log(\sum q^\alpha r^{1-\alpha})$.

	1% (0.3k)	**3%** (1.2k)	**25%** (9.6k)	**49%** (19.2k)	**97%** (38k)
random	70.61	77.21	82.98	84.48	85.51
w-js	74.07⋆	79.41⋆	83.98⋆	84.94⋆	85.68
w-var	74.07⋆	**79.60**⋆	83.82⋆	84.94⋆	85.45
w-skw	74.20⋆	78.95⋆	83.68⋆	84.60	85.55
w-cos	73.77⋆	79.30⋆	83.87⋆	84.96⋆	85.59
w-euc	73.85⋆	78.90⋆	83.52⋆	84.68	85.57
w-ryi	73.41⋆	78.31	83.76⋆	84.46	85.46
tm-js	74.23⋆	79.49⋆	84.04⋆	85.01⋆	85.45
tm-var	**74.29**⋆	79.59⋆	83.93⋆	84.94⋆	85.43
tm-skw	74.13⋆	79.42⋆	84.13⋆	84.82	**85.73**
tm-cos	74.04⋆	79.27⋆	**84.14**⋆	84.99⋆	85.42
tm-euc	74.27⋆	79.53⋆	83.93⋆	**85.15**⋆	85.62
tm-ryi	71.26	78.64⋆	83.79⋆	84.85	85.58

Table 2: Comparison of similarity measures based on words (w) and topic model (tm): accuracy for increasing amounts of training data as average over 22 WSJ articles (js=jensen-shannon; cos=cosine; skw=skew; var=variational; euc=euclidean; ryi=renyi). Best score (per representation) underlined, best overall score bold; ⋆ indicates significantly better ($p < 0.05$) than random.

Renyi divergence does not perform as well as other probabilistically-motivated functions. Regarding feature representations, the representation based on topic models works slightly better than the respective word-based measure (cf. Table 2) and often achieves the overall best score (boldface).

Overall, the differences in accuracy between the various similarity measures are small; but interestingly, the overlap between them is not that large. Table 3 and Table 4 show the overlap (in terms of proportion of identically selected articles) between pairs of similarity measures. As shown in Table 3, for all measures there is only a small overlap with the random baseline (around 10%-14%). Despite similar performance, topic model selection has interestingly no substantial overlap with any other word-based similarity measures: their overlap is at most 41.6%. Moreover, Table 4 compares the overlap of the various similarity functions within a certain feature representation (here x stands for either topic model – left value – or words – right value). The table shows that there is quite some overlap between jensen-shannon, variational and skew divergence on one side, and cosine and euclidean on the other side, i.e. between probabilistically- and geometrically-motivated functions. Variational has a higher overlap with the probabilistic functions. Interestingly, the 'peaks' in Table 4 (underlined, i.e. the highest pair-wise overlaps) are the same for the different feature representations.

In the following we analyze selection by topic model and words, as they are relatively different from each other, despite similar performance. For the word-based model, we use jensen-shannon as similarity function, as it turned out to be the best measure. For topic model, we use the simpler variational metric. However, very similar results were achieved using jensen-shannon. Cosine and euclidean did not perform as well.

	ran	w-js	w-var	w-skw	w-cos	w-euc
ran	–	10.3	10.4	10.0	10.4	10.2
tm-js	12.1	41.6	39.6	36.0	29.3	28.6
tm-var	12.3	40.8	39.3	34.9	29.3	28.5
tm-skw	11.8	40.9	39.7	36.8	30.0	30.1
tm-cos	14.0	31.7	30.7	27.3	24.1	23.2
tm-euc	14.6	27.5	27.2	23.4	22.6	22.1

Table 3: Average overlap (in %) of similarity measure: random selection (ran) vs. measures based on words (w) and topic model (tm).

x=tm/w	x-js	x-var	x-skw	x-cos	x-euc
tm/w-var	76/74	–	60/63	55/48	49/47
tm/w-skw	69/72	60/63	–	48/41	42/42
tm/w-cos	57/42	55/48	48/41	–	62/71
tm/w-euc	47/41	49/47	42/42	62/71	–

Table 4: Average overlap (in %) for different feature representations x as tm/w, where tm=topic model and w=words. Highest pair-wise overlap is underlined.

Automatic Measure vs. Human labels The next question is how these automatic measures compare to human-annotated data. We compare word-based and topic model selection (by using jensen-shannon and variational, respectively) to selection based on human-given labels: genre and topic. For genre, we randomly select larger amounts of training data for a given test article from the same genre. For topic, the approach is similar, but as an article might have several topic markers (keywords in the IN field), we

rank articles by proportion of overlapping topic keywords.

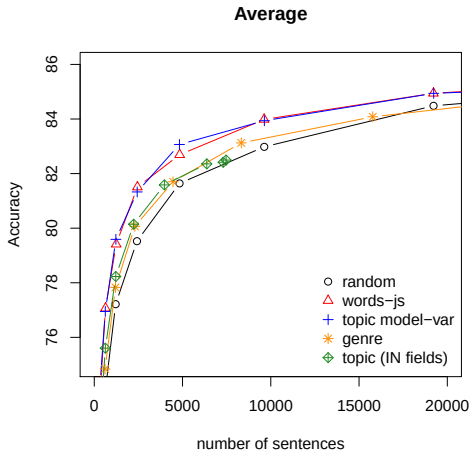

Figure 2: Comparison of automatic measures (words using jensen-shannon and topic model using variational) with human-annotated labels (genre/topic). Automatic measures outperform human labels ($p < 0.05$).

Figure 2 shows that human-labels do actually not perform better than the automatic measures. Both are just close to random selection. Moreover, the line of selection by topic marker (IN fields) stops early – we believe the reason for this is that the IN fields are too fine-grained, which limits the amount of articles that are considered relevant for a given test article. However, manually aggregating articles on similar topics did not improve topic-based selection either. We conclude that the automatic selection techniques perform significantly better than human-annotated data, at least within the WSJ domain considered here.

5.2 Domain Adaptation Results

Until now, we compared similarity measures by restricting ourselves to articles from the WSJ. In this section, we extend the experiments to the domain adaptation scenario. We augment the pool of WSJ articles with articles coming from two other corpora: Genia and Brown. We want to gauge the effectiveness of the domain similarity measures in the multi-domain setting, where articles are selected from the pool of data without knowing their identity (which corpus the articles came from).

The test sets are the standard evaluation sets from the three corpora: the standard WSJ (section 23) and Brown test set from CoNLL 2008 (they contain 2,399 and 426 sentences, respectively) and the Genia test set (1,370 sentences). As a reference, we give results of models trained on the respective corpora (*per-corpus* models; i.e. if we consider corpora boundaries and train a model on the respective domain – this model is 'supervised' in the sense that it knows from which corpus the test article came from) as well as a baseline model trained on all data, i.e. the *union* of all three corpora (wsj+genia+brown), which is a standard baseline in domain adaptation (Daumé III, 2007; McClosky et al., 2010).

	WSJ (38k)	**Brown** (28k)	**Genia** (19k)
random	86.58	73.81	83.77
per-corpus	87.50	81.55	86.63
union	87.05	79.12	81.57
topic model (var)	87.11⋆	81.76♢	86.77♢
words (js)	86.30	81.47♢	86.44♢

Table 5: Domain Adaptation Results on English (significantly better: ⋆ than random; ♢ than random and union).

The learning curves are shown in Figure 3, the scores for a specific amount of data are given in Table 5. The performance of the reference models (per-corpus and union in Table 5) are indicated in Figure 3 with horizontal lines: the dashed line represents the per-corpus performance ('supervised' model); the solid line shows the performance of the union baseline trained on all available data (77k sentences). For the former, the vertical dashed lines indicate the amount of data the model was trained on (e.g. 23k sentences for Brown).

Simply taking all available data has a deteriorating effect: on all three test sets, the performance of the union model is below the presumably best performance of a model trained on the respective corpus (per-corpus model).

The empirical results show that automatic data selection by topic model outperforms random selection on all three test sets and the union baseline in two out of three cases. More specifically, selection by topic model outperforms random selection significantly on all three test sets and all points in the graph ($p < 0.001$). Selection by the word-based measure (words-js) achieves a significant improve-

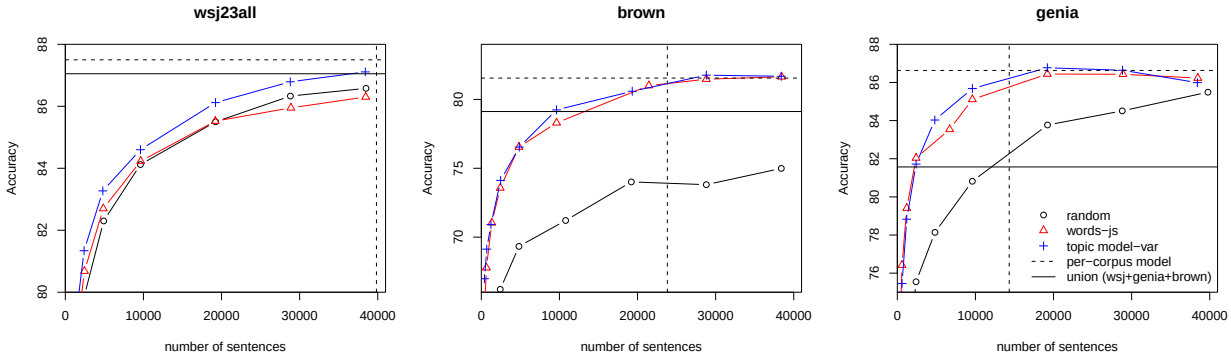

Figure 3: Domain Adaptation Results for English Parsing with Increasing Amounts of Training Data.

ment over the random baseline on two out of the three test sets – it falls below the random baseline on the WSJ test set. Thus, selection by topic model performs best – it achieves better performance than the union baseline with comparably little data (Genia: 4k; Brown: 19k – in comparison: union has 77k). Moreover, it comes very close to the supervised per-corpus model performance[13] with a similar amount of data (cf. vertical dashed line). This is a very good result, given that the technique disregards the origin of the articles and just uses plain words as information. It automatically finds data that is beneficial for an unknown target domain.

So far we examined domain similarity measures for parsing, and concluded that selection by topic model performs best, closely followed by word-based selection using the jensen-shannon divergence. The question that remains is whether the measure is more widely applicable: How does it perform on another language and task?

PoS tagging We perform similar Domain Adaptation experiments on WSJ, Genia and Brown for PoS tagging. We use two taggers (HMM and MaxEnt) and the same three test articles as before. The results are shown in Figure 4 (it depicts the average over the three test sets, WSJ, Genia, Brown, for space reasons). The left figure shows the performance of the HMM tagger; on the right is the MaxEnt tagger. The graphs show that automatic training data selection outperforms random data selection, and again topic model selection performs best, closely followed by words-js. This confirms previous findings and shows that the domain similarity measures are effective also for this task.

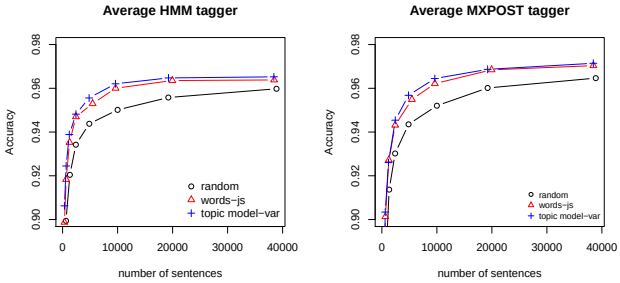

Figure 4: PoS tagging results, average over 3 test sets.

6 Experiments on Dutch

For Dutch, we evaluate the approach on a bigger and more varied dataset. It contains in total over 50k articles and 20 million words (cf. Table 1). In contrast to the English data, only a small portion of the dataset is manually annotated: 281 articles.[14]

Since we want to evaluate the performance of different similarity measures, we want to keep the influence of noise as low as possible. Therefore, we annotated the remaining articles with a parsing system that is more accurate (Plank and van Noord, 2010), the Alpino parser (van Noord, 2006). Note that using a more accurate parsing system to train another parser has recently also been proposed by Petrov et al. (2010) as *uptraining*. Alpino is a parser tailored to Dutch, that has been developed over the last ten years, and reaches an accuracy level

[13] On Genia and Brown (cf. Table 5) there is no significant difference between topic model and per-corpus model.

[14] http://www.let.rug.nl/vannoord/Lassy/

of 90% on general newspaper text. It uses a conditional MaxEnt model as parse selection component. Details of the parser are given in (van Noord, 2006).

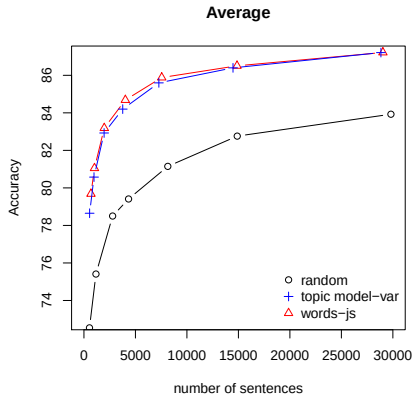

Figure 5: Result on Dutch; average over 30 articles.

Data and Results The Dutch dataset contains articles from a variety of sources: Wikipedia[15], EMEA[16] (documents from the European Medicines Agency) and the Dutch parallel corpus[17] (DPC), that covers a variety of subdomains. The Dutch articles were parsed with Alpino and automatically converted to CoNLL format with the treebank conversion software from CoNLL 2006, where PoS tags have been replaced with more fine-grained Alpino tags as that had a positive effect on MST. The 281 annotated articles come from all three sources. As with English, we consider as test set articles with at least 50 sentences, from which 30 are randomly sampled.

The results on Dutch are shown in Figure 5. Domain similarity measures clearly outperform random data selection also in this setting with another language and a considerably larger pool of data (20 million words; 51k articles).

7 Discussion

In this paper we have shown the effectiveness of a simple technique that considers only plain words as domain selection measure for two tasks, dependency parsing and PoS tagging. Interestingly, human-annotated labels did not perform better than the automatic measures. The best technique is based on topic models, and compares document topic distributions estimated by LDA (Blei et al., 2003) using the variational metric (very similar results were obtained using jensen-shannon). Topic model selection significantly outperforms random data selection on both examined languages, English and Dutch, and has a positive effect on PoS tagging. Moreover, it outperformed a standard Domain Adaptation baseline (union) on two out of three test sets. Topic model is closely followed by the word-based measure using jensen-shannon divergence. By examining the overlap between word-based and topic model-based techniques, we found that despite similar performance their overlap is rather small. Given these results and the fact that no optimization has been done for the topic model itself, results are encouraging: there might be an even better measure that exploits the information from both techniques. So far, we tested a simple combination of the two by selecting half of the articles by a measure based on words and the other half by a measure based on topic models (by testing different metrics). However, this simple combination technique did not improve results yet – topic model alone still performed best.

Overall, plain surface characteristics seem to carry important information of what kind of data is relevant for a given domain. Undoubtedly, parsing accuracy will be influenced by more factors than lexical information. Nevertheless, as we have seen, lexical differences constitute an important factor.

Applying divergence measures over syntactic patterns, adding additional articles to the pool of data (by uptraining (Petrov et al., 2010), selftraining (McClosky et al., 2006) or active learning (Hwa, 2004)), gauging the effect of weighting instances according to their similarity to the test data (Jiang and Zhai, 2007; Plank and Sima'an, 2008), as well as analyzing differences between gathered data are venues for further research.

Acknowledgments

The authors would like to thank Bonnie Webber and the three anonymous reviewers for their valuable comments on earlier drafts of this paper.

[15] http://ilps.science.uva.nl/WikiXML/
[16] http://urd.let.rug.nl/tiedeman/OPUS/EMEA.php
[17] http://www.kuleuven-kortrijk.be/DPC

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain Adaptation with Structural Correspondence Learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Meeting of the Association for Computational Linguistics*, Prague, Czech Republic.

Daniel Gildea. 2001. Corpus Variation and Parser Performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, Pittsburgh, PA.

Tadayoshi Hara, Yusuke Miyao, and Jun'ichi Tsujii. 2005. Adapting a Probabilistic Disambiguation Model of an HPSG Parser to a New Domain. In Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong, editors, *Natural Language Processing IJCNLP 2005*, volume 3651 of *Lecture Notes in Computer Science*, pages 199–210. Springer Berlin / Heidelberg.

Rebecca Hwa. 2004. Sample Selection for Statistical Parsing. *Compututational Linguistics*, 30:253–276, September.

Jing Jiang and ChengXiang Zhai. 2007. Instance Weighting for Domain Adaptation in NLP. In *Proceedings of the 45th Meeting of the Association for Computational Linguistics*, pages 264–271, Prague, Czech Republic, June. Association for Computational Linguistics.

Richard Johansson and Pierre Nugues. 2007. Extended Constituent-to-dependency Conversion for English. In *Proceedings of NODALIDA*, Tartu, Estonia.

Lillian Lee. 2001. On the Effectiveness of the Skew Divergence for Statistical Language Analysis. In *In Artificial Intelligence and Statistics 2001*, pages 65–72, Key West, Florida.

J. Lin. 1991. Divergence measures based on the Shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145 –151, January.

Tom Lippincott, Diarmuid Ó Séaghdha, Lin Sun, and Anna Korhonen. 2010. Exploring variation across biomedical subdomains. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 689–697, Beijing, China, August.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge Mass.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective Self-Training for Parsing. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 152–159, Brooklyn, New York. Association for Computational Linguistics.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic Domain Adaptation for Parsing. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36, Los Angeles, California, June. Association for Computational Linguistics.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden, July. Association for Computational Linguistics.

Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley-Interscience.

Slav Petrov, Pi-Chuan Chang, Michael Ringgaard, and Hiyan Alshawi. 2010. Uptraining for Accurate Deterministic Question Parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 705–713, Cambridge, MA, October. Association for Computational Linguistics.

Barbara Plank and Khalil Sima'an. 2008. Subdomain Sensitive Statistical Parsing using Raw Corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May.

Barbara Plank and Gertjan van Noord. 2010. Grammar-Driven versus Data-Driven: Which Parsing System Is More Affected by Domain Shifts? In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 25–33, Uppsala, Sweden, July. Association for Computational Linguistics.

Sujith Ravi, Kevin Knight, and Radu Soricut. 2008. Automatic Prediction of Parser Accuracy. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 887–

896, Morristown, NJ, USA. Association for Computational Linguistics.

A. Rényi. 1961. On measures of information and entropy. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, Berkeley.

Satoshi Sekine. 1997. The Domain Dependence of Parsing. In *In Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 96–102, Washington D.C.

Mark Steyvers and Tom Griffiths, 2007. *Probabilistic Topic Models*. Lawrence Erlbaum Associates.

Mihai Surdeanu and Christopher D. Manning. 2010. Ensemble Models for Dependency Parsing: Cheap and Good? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 649–652, Los Angeles, California, June. Association for Computational Linguistics.

Vincent Van Asch and Walter Daelemans. 2010. Using Domain Similarity for Performance Estimation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 31–36, Uppsala, Sweden, July. Association for Computational Linguistics.

Gertjan van Noord. 2006. At Last Parsing Is Now Operational. In *TALN 2006 Verbum Ex Machina, Actes De La 13e Conference sur Le Traitement Automatique des Langues naturelles*, pages 20–42, Leuven.

Bonnie Webber. 2009. Genre distinctions for Discourse in the Penn TreeBank. In *Proceedings of the 47th Meeting of the Association for Computational Linguistics*, pages 674–682, Suntec, Singapore, August. Association for Computational Linguistics.

Zhili Wu, Katja Markert, and Serge Sharoff. 2010. Fine-Grained Genre Classification Using Structural Learning Algorithms. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 749–759, Uppsala, Sweden, July. Association for Computational Linguistics.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics*, pages 947–953, Morristown, NJ, USA. Association for Computational Linguistics.