Knowledge-Poor Approach to Shallow Parsing: Contribution of Unsupervised Part-of-Speech Induction

Marie Guégan
Syllabs
15, rue Jean-Baptiste Berlier
75013 Paris, France
guegan@syllabs.com

Claude de Loupy
Syllabs
15, rue Jean-Baptiste Berlier
75013 Paris, France
loupy@syllabs.com

Abstract

Natural language processing tasks often rely on part-of-speech (POS) tagging as a preprocessing step. However it is not clear how the absence of any part-of-speech tagger should hamper the development of other natural language processing tools. In this paper we investigate the contribution of fully unsupervised part-of-speech induction to a common natural language processing task. We focus on the supervised English shallow parsing task and compare systems relying either on POS induction, on POS tagging, or on lexical features only as a baseline. Our experiments on the English CoNLL'2000 dataset show a significant benefit from POS induction over the baseline, with performances close to those obtained with a traditional POS tagger. Results demonstrate a great potential of POS induction for shallow parsing which could be applied to resource-scarce languages.

1 Introduction

Shallow parsing is a specific type of phrase chunking which is often used for different Natural Language Processing (NLP) tasks like text mining or question answering. The goal of the task is to divide a text into syntactically related non-overlapping groups of words (Tjong Kim Sang and Buchholz, 2000). These include noun, verb, or adjective phrases. It usually requires a part-of-speech (POS) tagger and a training corpus annotated with shallow parsing tags.

Unfortunately, one is often constrained by the lack of resources, tools or language experts, for instance when dealing with resource-scarce languages. In particular, the elaboration of a POS tagger is a delicate issue. Without any linguistic expert, the only possible approaches are statistical. Training POS taggers requires the manual constitution of either a large annotated corpus or a large morphosyntactic lexicon. These resources are very costly, both in time and in terms of linguistic knowledge required from the annotator.

By contrast, we notice that the concept of shallow parsing is relatively easily understandable by native speakers, even if they are not linguists. Relative to POS tagging, its annotation does not require a prohibitive amount of time and effort[1]. This is especially the case when the full shallow parsing task is reduced to a certain chunk type, as noun phrases for instance. Hence we think the most difficult requirement for the task is the POS tagging preprocessing step.

This observation drew our attention to the following question: is the POS tagging step necessary to shallow parsing? In this paper we intend to show how shallow parsing may benefit from fully unsupervised POS induction methods, as an alternative to accurate POS tagging. Section 2 introduces related work. Despite the popularity of shallow parsing and POS induction, we found only one paper related to POS induction for shallow parsing. Section 3 describes the models, tools and corpora we used: an existing POS induction tool (Clark, 2003), an implementation of Conditional Random Fields (CRF++) and the CoNLL'2000 dataset. Experiments and results are presented in Section 4. POS induction greatly improves the baseline, with performances close to supervised POS tagging.

2 Related Work

Shallow parsing has become a common task in NLP. The originality of our method is to rely on part-of-speech induction rather than accurate POS tagging.

[1] The standard English shallow parsing corpus contains around 50 distinct POS tags and only 10 chunk types.

Proceedings of Recent Advances in Natural Language Processing, pages 33–40,
Hissar, Bulgaria, 12-14 September 2011.

2.1 Shallow Parsing

Traditional approaches rely on preprocessing by an accurate POS tagger. Most work on shallow parsing is based on the English CoNLL'2000 shared task, which provided reference datasets for training and testing. The CoNLL dataset actually contains POS tags assigned by the Brill (1995) tagger. A number of approaches have been evaluated on these datasets, for general shallow parsing as well as for the simpler noun phrase chunking task: support vector machines (SVM) with polynomial kernels (Kudo and Matsumoto, 2001; Goldberg and Elhadad, 2009) and linear kernels (Lee and Wu, 2007), conditional random fields (Sha and Pereira, 2003), maximum likelihood trigram models (Shen and Sarkar, 2005), probabilistic finite-state automata (Araujo and Serrano, 2008), transformation-based learning or memory-based learning (Tjong Kim Sang, 2000). So far, SVM have achieved the best state-of-the-art performances.

To our knowledge, little work has considered other languages. Chunking corpora have been derived from the Arabic Treebank (Diab *et al.*, 2004) and the UPENN Chinese Treebank-4 (Chen *et al.*, 2006). Goldberg *et al.* (2006) showed that the traditional definition of base noun phrases as non-recursive noun phrases does not apply in Hebrew, and proposed an alternate definition. Nguyen *et al*. (2009) discuss on how to build annotated data for Vietnamese text chunking and how to apply discriminative sequence learning to Vietnamese text chunking. The lack of tools and annotated corpora in non-English languages is clearly an issue.

Following this observation and contrary to the approaches cited above, we make the assumption that no POS tagger is available. To compare our work with previous approaches and to allow extensive experiments, we evaluated our method on English using the standard CoNLL'2000 dataset. The lack of similar annotated corpora in other languages unfortunately constrained the scope of this article to English.

2.2 Part-of-Speech Induction

Unlike van den Bosch and Buchholz (2002) who studied shallow parsing on the basis of lexical features only, we choose to incorporate features related to the traditional notion of part of speech. In this work we apply part-of-speech induction techniques to acquire additional features. This task differs from semi-supervised part-of-speech tagging, where the tagger is trained on an untagged corpus but uses a morphosyntactic lexicon giving possible tags for each word (e.g. (Merialdo, 1994)). Part-of-speech induction is the task of clustering words into word classes (or *pseudo-POS*) in a completely unsupervised setting. No prior knowledge such as a morphosyntactic lexicon is required. The only resource needed is a relatively large training text corpus.

Christodoulopoulos *et al.* (2010) and (Biemann, 2010) compiled helpful surveys of the domain. Christodoulopoulos *et al.* (2010) evaluated seven POS induction systems spanning nearly 20 years of work: class-based n-grams (Brown *et al.*, 1992), class-based n-grams with morphology (Clark, 2003), Chinese Whispers graph clustering (Biemann, 2006), Bayesian HMM with Gibbs sampling (Goldwater and Griffiths, 2007), Bayesian HMM with variational Bayes (Johnson, 2007), sparsity posterior-regularization HMM (Graça *et al.*, 2009), and feature-based HMM (Berg-Kirkpatrick *et al.*, 2010). The performance measures were mainly based on mapping accuracies (with respect to a gold standard) and entropy coefficients.

Biemann *et al.* (2007) and Biemann (2010) succinctly tested their Chinese Whispers algorithm on the shallow parsing task with the English CoNLL'2000 dataset. They showed a significant improvement of the use of unsupervised pseudo part-of-speech tags over the baseline that discarded any POS information. However, their experiments covered several tasks and were not focused on shallow parsing. By contrast, in this article we use an alternate POS induction algorithm and propose a more in-depth evaluation of shallow parsing with POS induction.

3 Resources, Models and Tools

This section describes the tools and resources used in this work. Figure 1 depicts the global organization of our modules.

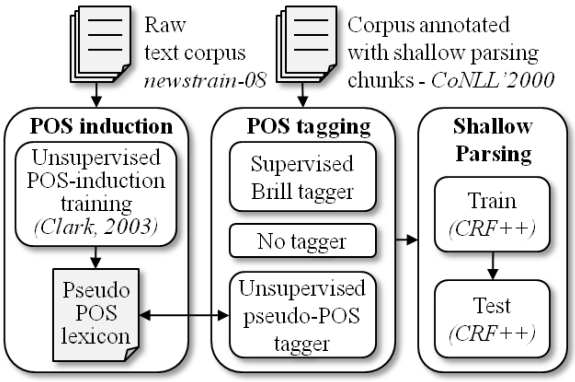

Figure 1. Overview of the system

On the left side of the figure, an unsupervised pseudo-POS tagger is learnt using POS induction techniques. This step requires a raw text corpus as input and produces a list of (word, cluster identifier) pairs which constitute the pseudo-POS lexicon. On the center, a POS tagger is optionally applied to the training and test corpora annotated with shallow parsing tags. Eventually, a supervised training of shallow parsing is conducted on the training set and evaluated on the test set (on the right).

The following sections describe the tools and corpora we used for the POS induction step and for the shallow parsing step. This information is summarized in Table 1.

3.1 Unsupervised POS Induction

Model and Tool

Based on Christodoulopoulos *et al.* (2010), we opted for Clark (2003)'s tool[2]. It was the best performing system in almost every language, and one of the fastest methods. It incorporates morphological information into a distributional clustering algorithm. To our knowledge, it has not yet been evaluated on the shallow parsing task.

The clustering algorithm is based on a cluster bigram model (Ney *et al.*, 1994). Assume we have of corpus of size N, composed of words $w_1 \dots w_N$. We note w_i^j the sequence of all words between i and j. We define a clustering function c that deterministically assigns a unique cluster identifier to each word form. The bigram model is a specific type of first-order hidden Markov model where each observation type (word form) is allowed to a single latent class. The model defines the probability of word w_i given history w_1^{i-1} and clustering c as:

$$P(w_i|w_1^{i-1}, c) = P(w_i|c(w_i)).P(c(w_i)|c(w_{i-1}))$$

In our case, the deterministic nature of the clustering makes the likelihood of the model easy to express in terms of word and cluster occurrence counts in the corpus given the clustering. The likelihood is maximized using an exchange algorithm similar to the *k*-means algorithm. It converges locally until a stopping criterion is reached. It consists in iteratively increasing the likelihood of an initial clustering by moving words one after the other to better clusters.

The morphological component biases the clustering so as to cluster together morphologically similar words. Clark (2003) models the morphology of words belonging to a same cluster using letter Hidden Markov Models and uses it to define a prior for this cluster in the basic cluster bigram model. The final output consists of a large table giving a unique cluster identifier to each word token, followed by the conditional probability of the word given the cluster. The pseudo POS tagging itself hence comes down to a simple deterministic look-up into the table.

Unlike Biemann (2010), the number of pseudo-POS clusters should be provided as a parameter of the algorithm. In our experiments, we learnt several pseudo-POS taggers with a number of clusters varying from 10 to 200 (see Section 4.3). Another parameter for Clark's tool is the token cutoff frequency. This threshold assigns all words occurring less than the specified number of times to a particular cluster. This cluster is the one that will be used for tagging unknown words.

Corpus

The tool takes a tokenized corpus as input. The corpus chosen for our experiments is *newstrain-08*, an English monolingual language model training dataset which was provided for the WMT'09 translation task[3]. Its size is approximately 2.5 Gb and 500 million tokens. We set the token cutoff frequency to 50[4].

Such enormous corpora might not be available for some languages. However we believe that the approach remains valid on smaller corpora. We therefore experimented on a subset of the newstrain-08 corpus restricted to the first million tokens only. To avoid losing too much information, the cutoff frequency was then set to 1: only hapaxes were discarded.

Step	Tool	Corpus	Corpus Size
POS induction	*(Clark, 2003)*	newstrain-08 full	500M tokens
		newstrain-08 short	1M tokens
Shallow Parsing	CRF++	CoNLL'2000 train	211,727 tokens 8936 sentences
		CoNLL'2000 test	47,377 tokens 2012 sentences

Table 1. Tools and corpora used for POS induction and shallow parsing

[2] Available on Alexander Clark's Web page: http://www.cs.rhul.ac.uk/home/alexc/pos2.tar.gz

[3] The corpus is available at: http://statmt.org/wmt09/training-monolingual.tar

[4] Other parameter values are "-s 5" (number of HMM states) and "-i 20" (stopping criterion: maximum number of iterations)

3.2 CRFs for Shallow Parsing

Model and Tool

We follow Sha and Pereira (2003), who achieved near state-of-the-art results on the English shallow parsing task using Conditional Random Fields (CRFs) (Lafferty *et al.*, 2003). CRFs allow us to incorporate a large number of features in a flexible way. We used the CRF++ implementation[5], distributed under the GNU Lesser General Public License and new BSD License.

Our feature set is defined as follows. On a 5-token window centered on the current token to be classified, we included all lowercased form token unigrams and bigrams, as well as (pseudo) POS tag unigrams, bigrams and trigrams. We also incorporated phrase chunk label bigrams. These features are commonly used for shallow parsing. Finally, we added on the same 5-token window a feature indicating whether the forms begin with a capital, as well as features accounting for the form ending (3 characters) on a window of 3 tokens. The purpose of these features is to facilitate the classification of unknown words by incorporating morphological information into the model.

In some experiments (see Section 4.4), we tried several feature frequency cutoff values, varying from 1 occurrence in the training set to at least 100. The default is set to 1.

Corpus

The standard reference corpus for English shallow parsing is the CoNLL'2000 shared task dataset. The CoNLL dataset[6] was automatically derived from a subset of the Wall Street Journal (WSJ) portion of the Penn Treebank. It consists of partitions of the WSJ: sections 15-18 as training data (8936 sentences) and section 20 as test data (2012 sentences). It contains phrase boundaries in the IOB representation, as well as part-of-speech tags assigned by the Brill tagger[7]. The corpus contains 48 Brill tags.

A sentence extracted from the CoNLL training corpus is shown in Table 2. Here, chunk phrases are separated with horizontal dashed lines. Each chunk type has 2 types of chunk labels: prefix B indicates the beginning of the chunk phrase, and prefix I stands for *inside the chunk phrase*. Label O represents tokens that do not belong to any phrase.

Token	Brill Tag	Chunk Label
A.P.	NNP	B-NP
Green	NNP	I-NP
currently	RB	B-ADVP
has	VBZ	B-VP
2,664,098	CD	B-NP
shares	NNS	I-NP
outstanding	JJ	B-ADJP
.	.	O

Table 2. Example sentence from the CoNLL'2000 training corpus

In some experiments, we discarded all Brill tags. In our POS-induction-based experiments, we replaced them with pseudo-POS tags.

4 Experiments and Results

Our experiments have 4 goals: (i) estimate the gain of POS induction over a system that does not rely on any part-of-speech information; (ii) estimate performance variation depending on the size of the shallow parsing training corpus; (iii) study the influence of the number of pseudo-POS clusters; (iv) observe the system behavior with CRF feature pruning. Our results were evaluated using the Perl script provided by CoNLL[8].

4.1 The CoNLL Shallow Parsing Task

We first evaluated our system in the traditional setting. Our objective is to estimate the potential of POS induction for shallow parsing in the case where no POS tagger is available.

We conducted three runs using the same CRF feature template (Section 3.2), depending on whether the POS tags are the original Brill tags from the corpus (*Brill*), our pseudo-POS tags (*P50*), or no tag at all as a baseline (*NoPOS*). For this experiment, we used the CoNLL datasets for training and testing. The pseudo-POS tagger was learnt on the full newstrain-08 corpus. We set the number of pseudo-POS tags to 50, which is comparable to the number of Brill tags.

Detailed results are presented in Table 3. It shows precision, recall and F-measure for each chunk category. Precision p is the percentage of correct phrases over the total number of phrases annotated by the system. Recall r is the percentage of correct phrases over the total number of true phrases in the reference. The F-measure F_1 is defined as the harmonic mean of precision and recall[9].

[5] Available at http://crfpp.sourceforge.net/

[6] See http://www.clips.ua.ac.be/conll2000/chunking/

[7] The original manually annotated tags from WSJ were discarded in order to make the CoNLL task more realistic.

[8] http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt

[9] $F_\beta = \frac{(1+\beta^2).p.r}{\beta^2.p+r}$ with $\beta = 1$

	Baseline: NoPOS				Unsupervised: P50				Supervised: Brill			
Chunk types	***p***	***r***	***F*$_1$**	**found**	***p***	***r***	***F*$_1$**	**found**	***p***	***r***	***F*$_1$**	**found**
ADJP	80.00	59.36	68.15	325	79.05	68.04	73.13	377	81.27	76.26	78.68	411
ADVP	84.21	75.75	79.76	779	82.94	80.25	81.57	838	84.24	80.83	82.50	831
CONJP	38.46	55.56	45.45	13	55.56	55.56	55.56	9	50.00	55.56	52.63	10
INTJ	100	100	100	2	100	50.00	66.67	1	100	50.00	66.67	1
LST	0	0	0	0	0	0	0	0	0	0	0	0
NP	91.38	91.01	91.19	12372	94.09	93.87	93.98	12393	94.44	94.13	94.29	12381
PP	96.71	97.15	96.93	4833	96.47	98.17	97.31	4896	96.84	98.03	97.43	4870
PRT	76.79	81.13	78.9	112	77.68	82.08	79.82	112	78.85	77.36	78.10	104
SBAR	85.82	83.74	84.77	522	89.49	85.98	87.70	514	88.10	85.79	86.93	521
VP	91.64	90.58	91.10	4604	93.33	93.09	93.21	4646	93.78	94.18	93.98	4678
All	91.91	90.79	**91.34**	23562	93.61	93.35	**93.48**	23786	93.99	93.82	**93.90**	23807
Accuracy				94.61				95.84				96.12

Table 3. Detailed chunking results for the English shallow parsing task

Column *found* in Table 3 gives the total number of phrases annotated by the system (correct or incorrect). *Accuracy* is the percentage of correct guesses at the token level.

First, we recall that the state-of-the art system of Lee and Wu (2007) reached a 94.22% F-measure using Brill part-of-speech tags. Comparably, our system performs reasonably well when using the same tags (Brill: F-measure 93.90%), considering that it was not subject to any refinements. Without any POS information, the system already achieves a high F-measure (91.34%).

We observe a 2% overall gain of P50 (F-measure 93.48%) over NoPOS, and a drop from Brill inferior to 0.5%. P50 beats Brill on a few categories, although not substantially: conjunctions, particles, and subordinating conjunctions. Its performances are very close to Brill on the 4 most frequent chunk types: adverb phrases, noun phrases, prepositional phrases, and verb phrases. These results incidentally suggest a great potential of the approach for noun phrase chunking, for which state-of-the-art systems reach about 96.8% F-measure (Araujo and Serrano, 2008).

The category of adjective phrases is the most difficult. Although significantly improving the recall of NoPOS, the F-measure for P50 lies exactly between NoPOS and Brill.

Looking into the test corpus

We examined the output test corpus to explain the differences between the unsupervised approach (P50) and the supervised approach (Brill). The accuracies tell us that on 47,377 tokens, Brill correctly tagged 130 tokens more than P50. Looking specifically at the 2,808 tokens that were unknown to the pseudo-POS tagger, Brill correctly tagged 16 tokens more than P50. Unknown words thus only account for 12% of the 130-token advantage.

The major sources of disagreement on chunks are shown in Table 4. These account for more than half the 130-token difference. It shows for instance that in cases where Brill chose B-NP and P50 chose I-NP, Brill was correct for 15 tokens more than P50. We observe that P50 tends to annotate too long noun and verb phrases (P50: incorrect I-NP and I-VP). It also shows more difficulties finding the beginning of verb and adjective phrases (Brill: B-VP and B-ADJP).

Finally, we examined the Brill parts of speech of misclassified chunks on which Brill and P50 disagreed (see Table 5). P50 mostly fails on adjectives and adverbs. Yet it better classifies IN tokens (preposition, subordinating conjunction).

Brill	P50	Brill correct	P50 correct	Diff
B-NP	I-NP	79	64	15
B-VP	B-PP	21	9	12
B-ADJP	B-ADVP	17	6	11
B-ADJP	B-VP	11	0	11
B-ADVP	I-NP	11	2	9
B-VP	B-NP	26	17	9
B-VP	I-VP	20	11	9

Table 4. Disagreement between Brill and P50 on a few chunks

Brill POS	Brill correct	P50 correct	Diff
JJ	87	49	38
RB	74	46	28
TO	34	21	13
VBG	37	26	11
VB	22	13	9
CC	56	49	7
IN	39	53	-14

Table 5. Disagreement between Brill and P50 on a few parts of speech

4.2 Training Corpus Size

Corpora such as the CoNLL'2000 dataset are expensive to produce and not yet available for many languages. Therefore we were interested in the evolution of performances with the size of the training corpus. We repeated the experiments from previous section on corpus sizes ranging from 1% (approximately 90 sentences) to 100% (approximately 9000 sentences). All systems were tested on the CoNLL test set. Each experiment was run on 20 different splits of the training corpus (except for the full corpus).

In addition, we wanted to take into account the difficulty of compiling large monolingual corpora in some languages. We therefore also tested the method using a much smaller corpus for POS induction training. It contains a subset of 1 million words from the newstrain-08 corpus, as opposed to 500 million for the full corpus (see Section 3.1). In this experiment we also set the number of pseudo-POS clusters to 50.

Figure 2 shows the F-measures for varying sizes of the training corpus on the abscissa on a logarithmic scale. The four curves correspond to the following taggers: Brill, pseudo POS tagger trained on the full newstrain-08 corpus (P50), pseudo POS tagger trained on the smaller newstrain corpus (P50m), and no tagger (NoPOS). Each point denotes the mean of the 20 runs. To give an insight of the variation in F-measure across all runs, we added box plots on the P50 curve. Each box is centered on the median of the runs. Half the points lie between its lower and upper sides. The whiskers extend to the most extreme data point which is no more than 1.5 times the height of the box away from the box.

We observe a significant improvement of our POS-induction-based systems over the baseline (NoPOS), especially for smaller training corpora. For a 1% sample of the CoNLL corpus, the F-measures are approximately 65% only for the baseline (NoPOS), 78% for the unsupervised systems (P50 and P50m) and 83.5% in the supervised setting (Brill).

A 90% F-measure is achieved starting from 10% of the training corpus by Brill, and starting from 20% by P50. More generally, the unsupervised system needs a little more than twice as much annotated data as the supervised system to achieve a similar F-measure.

With less than 200 sentences (2% sample), the unsupervised system almost achieves 83% F-measure, which is only achieved by the baseline starting from 900 sentences (10% sample).

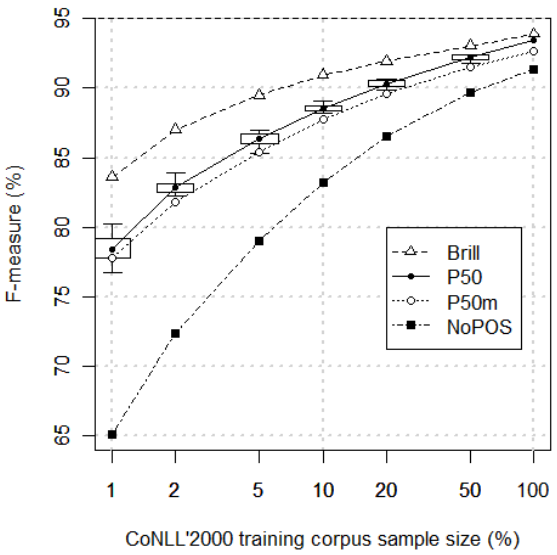

Figure 2. F-measure depending on the training corpus sample size and on the POS tagger

Finally, we notice that P50 and P50m get very close results, despite the fact that their pseudo POS taggers have been trained on 500M tokens and 1M tokens respectively. This result validates the approach for the case where only relatively small raw text corpora are available for training the pseudo POS tagger. This finding could be highly valuable for resource-scarce languages.

4.3 Number of pseudo-POS clusters

Some POS induction algorithms have the advantage over supervised POS tagging to easily adapt the number of word classes to the task.

Biemann (2010) conjectures for the same chunking task that results could be significantly improved with a smaller cluster number. To verify this hypothesis, we trained several pseudo-POS taggers with a cluster number between 10 and 200. Similarly to the experiments reported in the previous section, we evaluate the systems on varying sizes of the CoNLL training corpus[10].

Results are presented in Figure 3. From 10 to 50 clusters, performances increase with the number of clusters for all sizes of the training corpus. By contrast, P100 and P200 only improve over P50 for corpus sizes superior to 10%, which represents about 900 sentences. This can be attributed to the sparseness of pseudo POS tags in small training sets. We conclude that for small training corpus sizes, the number of pseudo-POS tags should be chosen carefully. On the whole, the F-measures vary in a 5.3% interval for a 1% sample, and in a 1.1% interval for the full corpus.

[10] Again, 20 runs for each size of the training corpus

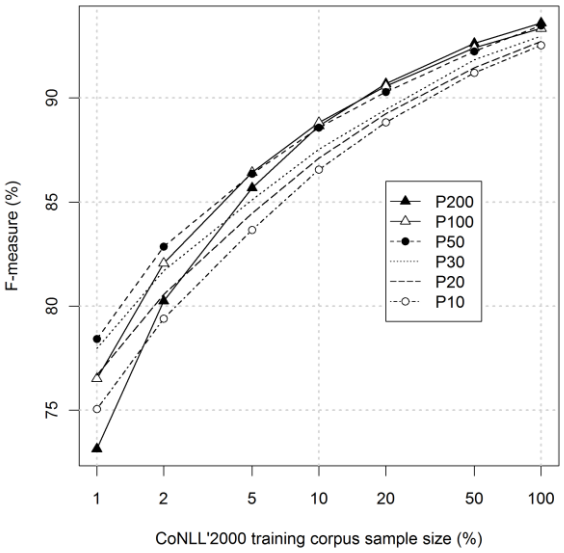

Figure 3. F-measure depending on the CoNLL'2000 training corpus sample size for a varying number of pseudo-POS clusters

The F-measures obtained using the full training dataset are: 93.6 (P200), 93.34 (P100), 93.48 (P50), 92.97 (P30), 92.72 (P20), and 92.53 (P10). They were 91.34 for the baseline and 93.9 for Brill (see Table 3): even 10 pseudo-POS clusters are sufficient to beat the baseline, and this is valid for all sizes of the training corpus.

4.4 CRF Feature Selection

In the last experiment we tested CRF feature pruning. The idea is to select the features appearing at least *k* times in the training corpus. This was motivated by Goldberg and Elhadad (2009), who explored the importance of lexical features in shallow parsing and other sequence labeling tasks. The performance of their anchored SVM system only decreased from 93.69% to 93.12% with heavy pruning ($k = 100$), while the baseline dropped from 93.73% to 91.83%. In addition, they showed comparable performances between heavily pruned models and full models when tested on out-of-domain data.

As in Goldberg and Elhadad (2009), we set the feature frequency threshold to values ranging from 1 to 100. Each experiment was run only once using the whole CoNLL training corpus.

Figure 4 shows that the supervised part-of-speech tagging system is the most robust to feature pruning. It loses less than 1% for $k = 100$. In comparison, the baseline NoPOS loses 4.3%. This indicates a strong dependency to the domain of the training corpus.

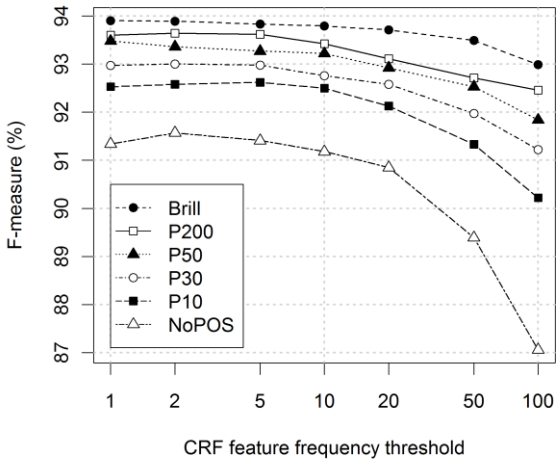

Figure 4. F-measure for various CRF feature pruning thresholds

The unsupervised systems resist quite well to feature pruning for $k < 20$, losing 1.1% and 1.6% F-measure for 200 and 50 clusters. P50 models have around 350,000 features for $k = 1$ and 5,100 features only for $k = 100$, while the baseline keeps from 270,000 to 2,200 features.

As in Goldberg and Elhadad (2009), it will be interesting to test the pruned models on out-of-domain corpora, and see how POS induction-based systems behave in comparison to systems relying on accurate part-of-speech information.

5 Conclusion and Future Work

In this paper, we study the contribution of part-of-speech induction to shallow parsing. The general context of our work is the automatic treatment of minority languages for which few linguistic resources are available, though we experimented on English only. Our constraint is the lack of any POS tagger. The experiments were carried out on the standard English CoNLL'2000 dataset, which allowed extensive experiments and explicit comparison to related work. We used Clark (2003)'s tool for the POS induction step and CRF++ for the shallow parsing train and test steps. Results show a significant advantage of POS-induction-based systems over a baseline which uses lexical features only.

In the future, we intend to apply these techniques to both shallow parsing and noun phrase chunking for minority languages. This will require the constitution of annotated corpora for training and testing. This paper shows that, for English, a corpus of 1 M words for POS induction, as well as a few hundred annotated sentences are enough to obtain interesting performances. If this could be proved on other lan-

guages, it could be a very interesting point to manage NLP for resource-scarce languages.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n° 248005[11].

References

Araujo, L., & Serrano, J. I. (2008). Highly accurate error-driven method for noun phrase detection. *Pattern Recognition Letters*, *29*(4), 547-557.

Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J., & Klein, D. (2010). Painless unsupervised learning with features. *Proceedings of HLT-NAACL 2010* (pp. 582-590).

Biemann, C. (2006). Unsupervised Part-of-Speech Tagging Employing Efficient Graph Clustering. *Proceedings of ACL-CoLing 2006 - Student Research Workshop* (pp. 7-12).

Biemann, C. (2010). Unsupervised Part-of-Speech Tagging in the Large. *Research on Language and Computation*, *7*(2-4), 101-135.

Biemann, C., Giuliano, C., & Gliozzo, A. (2007). Unsupervised Part of Speech Tagging Supporting Supervised Methods. *Proceedings of RANLP-07*.

van den Bosch, A., & Buchholz, S. (2001). Shallow parsing on the basis of words only. *Proceedings of ACL'02* (p. 433).

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics*, *21*(4), 543-565.

Brown, P. F., DeSouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics*, *18*(4), 467-479.

Chen, W., Zhang, Y., & Isahara, H. (2006). An Empirical Study of Chinese Chunking. *Proceedings of COLING/ACL 2006 Poster Sessions* (pp. 97-104).

Christodoulopoulos, C., Goldwater, S., & Steedman, M. (2010). Two decades of unsupervised POS induction: how far have we come? *Proceedings of EMNLP 2010* (pp. 575-584).

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. *Proceedings of EACL 2003* (pp. 59-66).

Diab, M., Hacioglu, K., & Jurafsky, D. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. *Proceedings of HLT-NAACL 2004: Short Papers* (p. 149–152).

Goldberg, Y., Adler, M., & Elhadad, M. (2006). Noun phrase chunking in Hebrew: influence of lexical and morphological features. *Proceedings of ACL-CoLing 2006* (pp. 689-696).

Goldberg, Y., & Elhadad, M. (2009). On the role of lexical features in sequence labeling. *Proceedings of EMNLP 2009* (pp. 1142-1151).

Goldwater, S., & Griffiths, T. L. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. *Proceedings of ACL'07* (pp. 744-751).

Graça, J., Ganchev, K., Taskar, B., & Pereira, F. (2009). Posterior vs. Parameter Sparsity in Latent Variable Models. *Proc. of NIPS* (p. 664–672).

Johnson, M. (2007). Why Doesn't EM Find Good HMM POS-Taggers? *Proceedings of EMNLP-CoNLL 2007*.

Kudo, T., & Matsumoto, Y. (2001). Chunking with support vector machines. *Proceedings of NAACL 2001* (pp. 1-8).

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of ICML'01* (pp. 282–289)

Lee, Y.-S., & Wu, Y.-C. (2007). A robust multilingual portable phrase chunking system. *Expert Systems with Applications*, *33*(3), 590-599.

Merialdo, B. (1994). Tagging English text with a probabilistic model. *Computational Linguistics*, *20*(2), 155--171. MIT Press.

Ney, H., Essen, U., & Kneser, R. (1994). On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, *8*(1), 1-38.

Nguyen, L. M., Nguyen, H. T., Nguyen, P. T., Ho, T. B., & Shimazu, A. (2009). An empirical study of Vietnamese noun phrase chunking with discriminative sequence models. *Proc. of the 7th Workshop on Asian Language Resources* (pp. 9-16).

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. *Proceedings of NAACL 2003* (pp. 134-141).

Shen, H., & Sarkar, A. (2005). Voting Between Multiple Data Representations for Text Chunking. In *Advances in Artificial Intelligence* (Vol. 3501, pp. 389-400). Springer Berlin / Heidelberg.

Tjong Kim Sang, E. F. (2000). Noun Phrase Recognition by System Combination. *Proceedings of NAACL 2000* (p. 6).

Tjong Kim Sang, E. F., & Buchholz, S. (2000). Introduction to the CoNLL-2000 shared task. *Proceedings of CoNLL 2000 - LLL 2000* (pp. 127-132).

[11] http://www.ttc-project.eu