Wiki-ly Supervised Part-of-Speech Tagging

Shen Li
Computer & Information Science
University of Pennsylvania
shenli@seas.upenn.edu

João V. Graça
L^{2}F INESC-ID
Lisboa, Portugal
javg@l2f.inesc-id.pt

Ben Taskar
Computer & Information Science
University of Pennsylvania
taskar@cis.upenn.edu

Abstract

Despite significant recent work, purely unsupervised techniques for part-of-speech (POS) tagging have not achieved useful accuracies required by many language processing tasks. Use of parallel text between resource-rich and resource-poor languages is one source of weak supervision that significantly improves accuracy. However, parallel text is not always available and techniques for using it require multiple complex algorithmic steps. In this paper we show that we can build POS-taggers exceeding state-of-the-art bilingual methods by using simple hidden Markov models and a freely available and naturally growing resource, the Wiktionary. Across eight languages for which we have labeled data to evaluate results, we achieve accuracy that significantly exceeds best unsupervised and parallel text methods. We achieve highest accuracy reported for several languages and show that our approach yields better out-of-domain taggers than those trained using fully supervised Penn Treebank.

1 Introduction

Part-of-speech categories are elementary building blocks that play an important role in many natural language processing tasks, from machine translation to information extraction. Supervised learning of taggers from POS-annotated training text is a well-studied task, with several methods achieving near-human tagging accuracy (Ratnaparkhi, 1996; Toutanova et al., 2003; Shen et al., 2007). However, while English and a handful of other languages are fortunate enough to have comprehensive POS-annotated corpora such as the Penn Treebank (Marcus et al., 1993), most of the world's languages have no labeled corpora. The annotated corpora that do exist were costly to build (Abeillé, 2003), and are often not freely available or restricted to research-only use. Furthermore, much of the annotated text is of limited genre, normally focusing on newswire or literary text. Performance of treebank-trained systems degrades significantly when applied to new domains (Blitzer et al., 2006).

Unsupervised induction of POS taggers offers the possibility of avoiding costly annotation, but despite recent progress, the accuracy of unsupervised POS taggers still falls far behind supervised systems, and is not suitable for most applications (Berg-Kirkpatrick et al., 2010; Graça et al., 2011; Lee et al., 2010). Using additional information, in the form of tag dictionaries or parallel text, seems unavoidable at present. Early work on using tag dictionaries used a labeled corpus to extract all allowed word-tag pairs (Merialdo, 1994), which is quite an unrealistic scenario. More recent work has used a subset of the observed word-tag pairs and focused on generalizing dictionary entries (Smith and Eisner, 2005; Haghighi and Klein, 2006; Toutanova and Johnson, 2007; Goldwater and Griffiths, 2007). Using corpus-based dictionaries greatly biases the test results, and gives little information about the capacity to generalize to different domains.

Recent work by Das and Petrov (2011) builds a dictionary for a particular language by transferring annotated data from a resource-rich language through the use of word alignments in parallel text.

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1389–1398, Jeju Island, Korea, 12–14 July 2012.

The main idea is to rely on existing dictionaries for some languages (e.g. English) and use parallel data to build a dictionary in the desired language and extend the dictionary coverage using label propagation. However, parallel text does not exist for many pairs of languages and the proposed bilingual projection algorithms are fairly complex.

In this work we use the Wiktionary, a freely available, high coverage and constantly growing dictionary for a large number of languages. We experiment with a very simple second-order Hidden Markov Model with feature-based emissions (Berg-Kirkpatrick et al., 2010; Graça et al., 2011). We outperform best current results using parallel text supervision across 8 different languages, even when the word type coverage is as low as 20%. Furthermore, using the Brown corpus as out-of-domain data we show that using the Wiktionary produces better taggers than using the Penn Treebank dictionary (88.5% vs 85.9%). Our empirical analysis and the natural growth rate of the Wiktionary suggest that free, high-quality and multi-domain POS-taggers for a large number of languages can be obtained by standard and efficient models.

The source code, the dictionary mappings and the trained models described in this work are available at `http://code.google.com/p/wikily-supervised-pos-tagger/`.

2 Related Work

The scarcity of labeled corpora for resource poor languages and the challenges of domain adaptation have led to several efforts to build systems for unsupervised POStagging.

Several lines of research have addressed the fully unsupervised POS-tagging task: mutual information clustering (Brown et al., 1992; Clark, 2003) has been used to group words according to their distributional context. Using dimensionality reduction on word contexts followed by clustering has led to accuracy gains (Schütze, 1995; Lamar et al., 2010). Sequence models, HMMs in particular, have been used to represent the probabilistic dependencies between consecutive tags. In these approaches, each observation corresponds to a particular word and each hidden state corresponds to a cluster. However, using maximum likelihood training for such models does not achieve good results (Clark, 2003): maximum likelihood training tends to result in very ambiguous distributions for common words, in contradiction with the rather sparse word-tag distribution. Several approaches have been proposed to mitigate this problem, including Bayesian approaches using an improper Dirichlet prior to favor sparse model parameters (Johnson, 2007; Gao and Johnson, 2008; Goldwater and Griffiths, 2007), or using the Posterior Regularization to penalize ambiguous posteriors distributions of tags given tokens (Graça et al., 2009). Berg-Kirkpatrick et al. (2010) and Graça et al. (2011) proposed replacing the multinomial emission distributions of standard HMMs by maximum entropy (ME) feature-based distributions. This allows the use of features to capture morphological information, and achieves very promising results. Despite these improvements, fully unsupervised systems require an oracle to map clusters to true tags and the performance still fails to be of practical use.

In this paper we follow a different line of work where we rely on a prior tag dictionary indicating for each word type what POS tags it can take on (Merialdo, 1994). The task is then, for each word token in the corpus, to disambiguate between the possible POS tags. Even when using a tag dictionary, disambiguating from all possible tags is still a hard problem and the accuracy of these methods is still fall far behind their supervised counterparts. The scarcity of large, manually-constructed tag dictionaries led to the development of methods that try to generalize from a small dictionary with only a handful of entries (Smith and Eisner, 2005; Haghighi and Klein, 2006; Toutanova and Johnson, 2007; Goldwater and Griffiths, 2007), however most previous works build the dictionary from the labeled corpus they learn on, which does not represent a realistic dictionary. In this paper, we argue that the Wiktionary can serve as an effective and much less biased tag dictionary.

We note that most of the previous dictionary based approaches can be applied using the Wiktionary and would likely lead to similar accuracy increases that we show in this paper. For example, the work if Ravi and Knight (2009) minimizes the number of possible tag-tag transitions in the HMM via a integer program, hence discarding unlikely transitions that would confuse the model. Models can also be trained jointly using parallel corpora in sev-

eral languages, exploiting the fact that different languages present different ambiguities (Snyder et al., 2008).

The Wiktionary has been used extensively for other tasks such as domain specific information retrieval (Müller and Gurevych, 2009), ontology matching (Krizhanovsky and Lin, 2009), synonymy detection (Navarro et al., 2009), sentiment classification (Chesley et al., 2006). Recently, Ding (2011) used the Wiktionary to initialize an HMM for Chinese POS tagging combined with label propagation.

3 The Wiktionary and tagged corpora

The Wiktionary[1] is a collaborative project that aims to produce a free, large-scale multilingual dictionary. Its goal is to describe all words from all languages (currently more than 400) using definitions and descriptions in English. The coverage of the Wiktionary varies greatly between languages: currently there are around 75 languages for which there exists more than 1000 word types, and 27 for which there exists more than 10,000 word types. Nevertheless, the Wiktionary has been growing at a considerable rate (see Figure 1), and the number of available words has almost doubled in the last three years. As more people use the Wiktionary, it is likely to grow. Unlike tagged corpora, the Wiktionary provides natural incentives for users to contribute missing entries and expand this communal resource akin to Wikipedia. As with Wikipedia, the questions of accuracy, bias, consistency across languages, and selective coverage are paramount. In this section, we explore these concerns by comparing Wiktionary to dictionaries derived from tagged corpora.

3.1 Labeled corpora and Universal tags

We collected part-of-speech tagged corpora for 9 languages, from CoNLL-X and CoNLL-2007 shared tasks on dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007). In this work we use the Universal POS tag set (Petrov et al., 2011) that defines 12 universal categories with a relatively stable functional definition across languages. These categories include NOUN, VERB, ADJ = adjective, ADV = adverb, NUM = number, ADP = adposition, CONJ = conjunction, DET = determiner, PRON = pronoun, PUNC = punctuation, PRT = particle, and X = residual (a category for language-specific categories which defy cross-linguistic classification). We found several small problems with the mapping[2] which we corrected as follows. In Spanish, the fine-level tag for date ("w") is mapped to universal tag NUM, while it should be mapped to NOUN. In Danish there were no PRT, NUM, PUNC, or DET tags in the mapping. After examining the corpus guidelines and the mapping more closely, we found that the tag AC (Cardinal numeral) and AO (Ordinal numeral) are mapped to ADJ. Although the corpus guidelines indicate the category SsCatgram 'adjective' that encompasses both 'normal' adjectives (AN) as well as cardinal numeral (AC) and ordinal numerals (AO), we decided to tag AC and AO as NUM, since this assignment better fits the existing mapping. We also reassigned all punctuation marks, which were erroneously mapped to X, to PUNC and the tag U which is used for words *at, de* and *som*, to PRT.

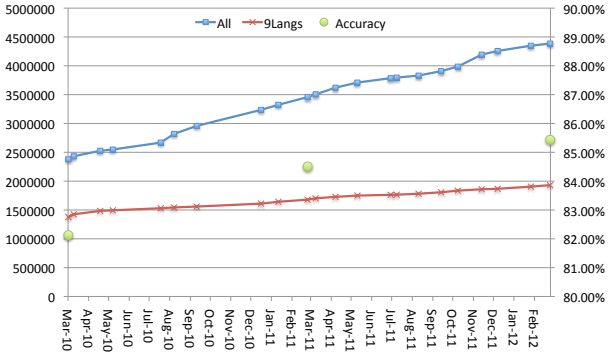

Figure 1: Growth of the Wiktionary over the last three years, showing total number of entries for all languages and for the 9 languages we consider (left axis). We also show the corresponding increase in average accuracy (right axis) achieved by our model across the 9 languages (see details below).

3.2 Wiktionary to Universal tags

There are a total of 330 distinct POS-type tags in Wiktionary across all languages which we have mapped to the Universal tagset. Most of the mapping was straightforward since the tags used in the Wiktionary are in fact close to the Universal tag set. Some exceptions like "Initialism", "Suffix"

[1] http://www.wiktionary.org/

[2] http://code.google.com/p/universal-pos-tags/

were discarded. We also mapped relatively rare tags such as "Interjection", "Symbol" to the "X" tag. A example of POS tags for several words in the Wiktionary is shown in Table 1. All the mappings are available at `http://code.google.com/p/wikily-supervised-pos-tagger/`.

3.3 Wiktionary coverage

There are two kinds of coverage of interest: type coverage and token coverage. We define type coverage as the proportion of word types in the corpus that simply appear in the Wiktionary (accuracy of the tag sets are considered in the next subsection). Token coverage is defined similarly as the portion of all word tokens in the corpus that appear in the Wiktionary. These statistics reflect two aspects of the usefulness of a dictionary that affect learning in different ways: token coverage increases the density of supervised signal while type coverage increases the diversity of word shape supervision. At one extreme, with 100% word and token coverage, we recover the POS tag disambiguation scenario and, on the other extreme of 0% coverage, we recover the unsupervised POS induction scenario.

The type and token coverage of Wiktionary for each of the languages we are using for evaluation is shown in Figure 2. We plot the coverage bar for three different versions of Wiktionary (v20100326, v20110321, v20120320), arranged chronologically. We chose these three versions of the Wiktionary simply by date, not any other factors like coverage, quality or tagging accuracy.

As expected, the newer versions of the Wiktionary generally have larger coverage both on type level and token level. Nevertheless, even for languages whose type coverage is relatively low, such as Greek (el), the token level coverage is still quite good (more than half of the tokens are covered). The reason for this is likely the bias of the contributors towards more frequent words. This trend is even more evident when we break up the coverage by frequency of the words. Since the number of words varies from corpus to corpus, we normalize the word counts by the count of the most frequent word(s) in its corpus and group the normalized frequency into three categories labeled as "low", "medium" and "high" and for each category, we calculate the word type coverage, shown in Figure 3.

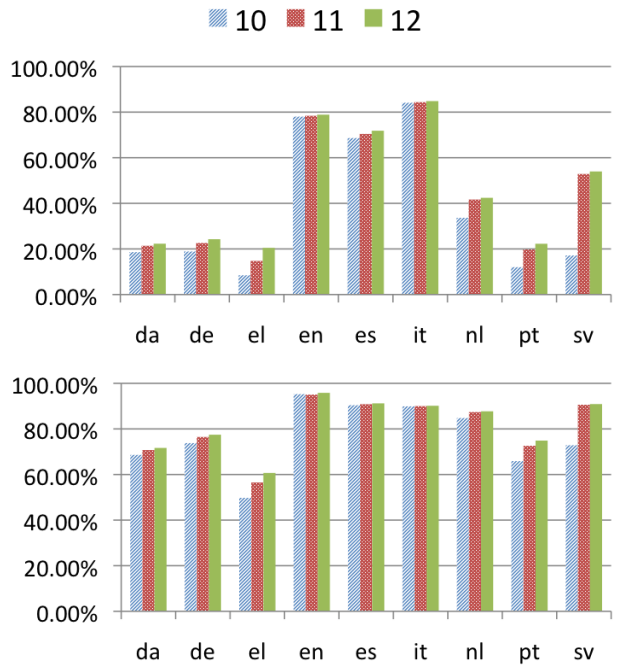

Figure 2: Type-level (top) and token-level (bottom) coverage for the nine languages in three versions of the Wiktionary.

We also compared the coverage provided by the Wiktionary versus the Penn Treebank (PTB) extracted dictionary on the Brown corpus. Figure 4 shows that the Wiktionary provides a greater coverage for all sections of the Brown corpus, hence being a better dictionary for tagging English text in general. This is also reflected in the gain in accuracy on Brown over the taggers learned from the PTB dictionary in our experiments.

3.4 Wiktionary accuracy

A more refined notion of quality is the accuracy of the tag sets for covered words, as measured against dictionaries extracted from labeled tree bank corpora. We consider word types that are in both the Wiktionary (W) and the tree bank dictionaries (T). For each word type, we compare the two tag sets and distinguish five different possibilities:

1. Identical: $W = T$
2. Superset: $W \supset T$
3. Subset: $W \subset T$
4. Overlap: $W \cap T \neq \emptyset$

Wiktionary Entries				Universal POS Set
Language	Word	POS	Definition	
English	today	Adverb	# In the current [[era]]; nowadays.	{ADV, NOUN}
English	today	Adverb	# On the current [[day]] or [[date]].	
English	today	Noun	# A current day or date.	
German	achtzig	Numeral	# [[eighty]]	{NUM}
Swedish	SCB	Acronym	# [[statistiska]] ...	{NOUN}
Portuguese	nessa	Contraction	# {{contraction ...	*discard entry*

Table 1: Examples of constructing Universal POS tag sets from the Wiktionary.

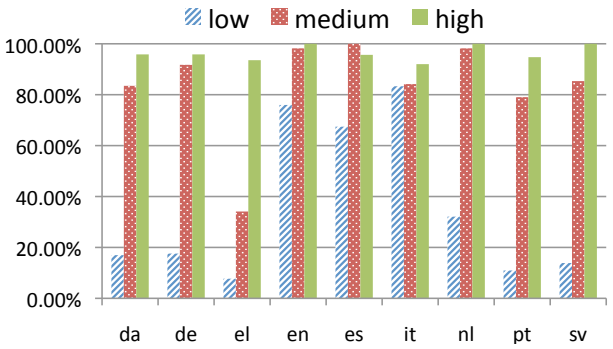

Figure 3: Word type coverage by normalized frequency: words are grouped by word count / highest word count ratio: low [0, 0.01), medium [0.01, 0.1), high [0.1, 1].

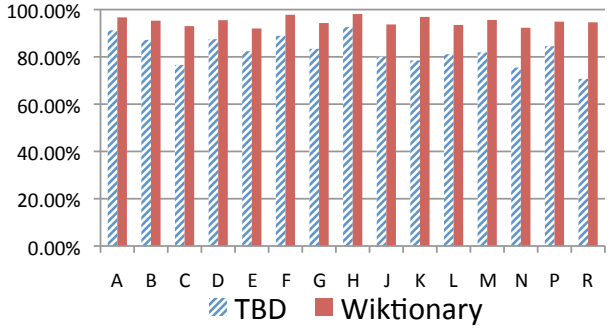

Figure 4: PTB vs. Wiktionary type coverage across sections of the Brown corpus.

5. Disjoint: $W \cap T = \emptyset$.

In Figure 5, the word types are grouped into the categories described above. Most of the tag sets (around 90%) in the Wiktionary are identical to or supersets of the tree bank tag sets for our nine languages, which is surprisingly accurate. About 10% of the Wiktionary tag sets are subsets of, partially overlapping with, or disjoint from the tree bank tag sets. Our learning methods, which assume the given tag sets are correct, may be somewhat hurt by these word types, as we discuss in Section 5.6.

4 Models

Our basic models are first and second order Hidden Markov Models (HMM and SHMM). We also used feature-based max-ent emission models with both (HMM-ME and SHMM-ME). Below, we denote the sequence of words in a sentence as boldface $\mathbf{x}$ and the sequence of hidden states which correspond to part-of-speech tags as boldface $\mathbf{y}$. To simplify notation, we assume that every tag sequence is prefixed with two conventional start tags $y_0 = \texttt{start}, y_{-1} = \texttt{start}$, allowing us to write as $p(y_1|y_0, y_{-1})$ the initial state probability of the SHMM.

The probability of a sentence $\mathbf{x}$ along with a particular hidden state sequence $\mathbf{y}$ in the SHMM is given by:

$$p(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{\text{length}(\mathbf{x})} p_t(y_i \mid y_{i-1}, y_{i-2}) p_o(x_i \mid y_i), \quad (1)$$

where $p_o(x_i \mid y_i)$ is the probability of observing word x_i in state y_i (emission probability), and $p_t(y_i \mid y_{i-1}, y_{i-2})$ is the probability of being in state y_i, given two previous states y_{i-1}, y_{i-2} (transition probability).

In this work, we compare multinomial and maximum entropy (log-linear) emission models. Specifically, the max-ent emission model is:

$$p_o(x|y) = \frac{\exp(\theta \cdot \mathbf{f}(x, y))}{\sum_{x'} \exp(\theta \cdot \mathbf{f}(x', y))} \quad (2)$$

where $\mathbf{f}(x, y)$ is a feature function, x ranges over all

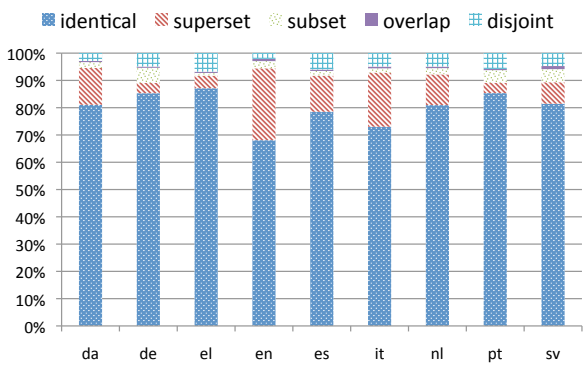

Figure 5: The Wiktionary vs. tree bank tag sets. Around 90% of the Wiktionary tag sets are identical or subsume tree bank tag sets. See text for details.

word types, and θ are the model parameters. We use the following feature templates:

- Word identity - lowercased word form if the word appears more than 10 times in the corpus.
- Hyphen - word contains a hyphen
- Capital - word is uppercased
- Suffix - last 2 and 3 letters of a word if they appear in more than 20 different word types.
- Number - word contains a digit

The idea of replacing the multinomial models of an HMM by maximum entropy models has been applied before in different domains (Chen, 2003), as well as in POS induction (Berg-Kirkpatrick et al., 2010; Graça et al., 2011).

We use the EM algorithm to learn the models, restricting the tags of each word to those specified by the dictionary. For each tag y, the observations probabilities $p_o(x \mid y)$ were initialized randomly for every word type that allows tag y according to the Wiktionary and zero otherwise. For the M-step in max-ent models, there is no closed form solution so we need to solve an unconstrained optimization problem. We use L-BFGS with Wolfe's rule line search (Nocedal and Wright, 1999). We found that EM achieved higher accuracy across languages compared to direct gradient approach (Berg-Kirkpatrick et al., 2010).

5 Results

We evaluate the accuracy of taggers trained using the Wiktionary using the 4 different models: A first order Hidden Markov Model (HMM), a second order Hidden Markov Model (SHMM), a first order Hidden Markov Model with Maximum Entropy emission models (HMM-ME) and a second order Hidden Markov Model with Maximum Entropy emission models (SHMM-ME). For each model we ran EM for 50 iterations, which was sufficient for convergence of the likelihood. Following previous work (Graça et al., 2011), we used a Gaussian prior with variance of 10 for the max-ent model parameters. We obtain hard assignments using posterior decoding, where for each position we pick the label with highest posterior probability: this produces small but consistent improvements over Viterbi decoding.

5.1 Upper and lower bounds

We situate our results against several upper bounds that use more supervision. We trained the SHMM-ME model with a dictionary built from the training and test tree bank (ALL TBD) and also with tree bank dictionary intersected with the Wiktionary (Covered TBD). The Covered TBD dictionary is more supervised than the Wiktionary in the sense that some of the tag set mismatches of the Wiktionary are cleaned using the true corpus tags. We also report results from training the SHMM-ME in the standard supervised fashion, using 50 (50 Sent.), 100 (100 Sent.) and all sentences (All Sent.).

As a lower bound we include the results for unsupervised systems: a regular HMM model trained with EM (Johnson, 2007) and an HMM model using a ME emission model trained using direct gradient (Berg-Kirkpatrick et al., 2010)[3].

5.2 Bilingual baselines

Finally, we also compare our system against a strong set of baselines that use bilingual data. These approaches build a dictionary by transferring labeled data from a resource rich language (English) to a resource poor language (Das and Petrov, 2011). We compare against two such methods. The first, *projection*, builds a dictionary by transferring the pos

[3] Values for these systems where taken from the D&P paper.

tags from English to the new language using word alignments. The second method, *D&P*, is the current state-of-the-art system, and runs label propagation on the dictionary resulting from the *projected* method. We note that both of these approaches are orthogonal to ours and could be used simultaneously with the Wiktionary.

5.3 Analysis

Table 2 shows results for the different models across languages. We note that the results are not directly comparable since both the Unsupervised and the Bilingual results use a different setup, using the number of fine grained tags for each language as hidden states instead of 12 (as we do). This greatly increases the degrees of freedom of the model allowing it to capture more fine grained distinctions.

The first two observations are that using the ME entropy emission model always improves over the standard multinomial model, and using a second order model always performs better. Comparing with the work of *D&P*, we see that our model achieves better accuracy on average and on 5 out of 8 languages.

The most common errors are due to tag set idiosyncrasies. For instance, for English the symbol % is tagged as NUM by our system while in the Penn treebank it is tagged as Noun. Other common mistakes for English include tagging *to* as an adposition (preposition) instead of particle and tagging *which* as a pronoun instead of determiner. In the next subsections we analyze the errors in more detail.

Finally, for English we also trained the SHMM-ME model using the Celex2 dictionary available from LDC[4]. Celex2 coverage for the PTB corpus is much smaller than the coverage provided by the Wiktionary (43.8% type coverage versus 80.0%). Correspondingly, the accuracy of the model trained using Celex2 is 75.5% compared 87.1% when trained using the Wiktionary.

5.4 Performance vs. Wiktionary ambiguity

While many words overwhelmingly appear with one tag in a given genre, in the Wiktionary a large proportion of words are annotated with several tags, even when those are extremely rare events. Around 35% of word types in English have more than one tag according to the Wiktionary. This increases the difficulty of predicting the correct tag as compared to having a corpus-based dictionary, where words have a smaller level of ambiguity. For example, in English, for words with one tag, the accuracy is 95% (the reason it is not 100% is due to a discrepancy between the Wiktionary and the tree bank.) For words with two possible tags, accuracy is 81% and for three tags, it drops to 63%.

5.5 Generalization to unknown words

Comparing the performance of the proposed model for words in the Wiktionary against words not in the Wiktionary, we see an average drop from 89% to 63% for out-of-vocabulary words across nine languages. Table 2 shows that the average loss of accuracy between All TBD and Covered TBD of 4.5% (which is due purely to decrease in coverage) is larger than the loss between Covered TBD and the best Wiktionary model, of 3.2% (which is due to tag set inconsistency).

One advantage of the Wiktionary is that it is a general purpose dictionary and not tailored for a particular domain. To illustrate this we compared several models on the Brown corpus: the SHMM-ME model using the Wiktionary (Wik), against using a model trained using a dictionary extracted from the PTB corpus (PTBD), or trained fully supervised using the PTB corpus (PTB). We tested all these models on the 15 different sections of the Brown corpus. We also compare against a state-of-the-art POS-tagger tagger (ST)[5].

Figure 6 shows the accuracy results for each model on the different sections. The fully supervised SHMM-ME model did not perform as well as the the Stanford tagger (about 3% behind on average), most likely because of generative vs. discriminate training of the two models and feature differences. However, quite surprisingly, the Wiktionary-tag-set-trained model performs much better not only than the PTB-tag-set-trained model but also the supervised model on the Brown corpus.

[4] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC96L14

[5] Available at http://nlp.stanford.edu/software/tagger.shtml

		Danish	Dutch	German	Greek	English	Italian	Portuguese	Spanish	Swedish	*avg.*
Unsupervised	HMM	68.7	57.0	75.9	65.8		63.7	62.9	71.5	68.4	*66.7*
	HMM-ME	69.1	65.1	81.3	71.8		68.1	78.4	80.2	70.1	*73.0*
Bilingual	Projection	73.6	77.0	83.2	79.3		79.7	82.6	80.1	74.7	*78.8*
	D&P	83.2	79.5	82.8	**82.5**		**86.8**	**87.9**	84.2	80.5	*83.4*
Wiktionary	HMM	71.8	80.8	77.1	73.1	85.4	84.6	79.1	83.9	76.7	*78.4*
	HMM-ME	82.8	86.1	81.2	80.1	86.1	85.4	83.7	84.6	85.9	*83.7*
	SHMM	74.5	81.6	81.2	73.1	85.0	85.2	79.9	84.5	78.7	*79.8*
	SHMM-ME	**83.3**	**86.3**	**85.8**	79.2	**87.1**	86.5	84.5	**86.4**	**86.1**	***84.8***
Supervised	Covered TBD	90.1	91.4	89.4	79.7	92.7	86.3	91.5	85.1	91.0	*88.6*
	All TBD	93.6	91.2	95.6	87.9	90.6	92.9	91.2	92.1	83.8	*91.0*
	50 Sent.	65.3	48.5	74.5	74.2	70.2	76.2	79.2	76.2	54.7	*68.6*
	100 Sent.	73.9	52.3	80.9	81.6	77.3	75.3	82.0	80.1	64.8	*73.9*
	All Sent.	93.9	90.9	97.4	95.1	95.8	93.8	95.5	93.8	95.5	*94.5*

Table 2: Accuracy for Unsupervised, Bilingual, Wiktionary and Supervised models. Avg. is the average of all languages except English. Unsupervised models are trained without dictionary and use an oracle to map tags to clusters. Bilingual systems are trained using a dictionary transferred from English into the target language using word alignments. The *Projection* model uses a dictionary build directly from the part-of-speech projection. The *D&P* model extends the *Projection* model dictionary by using Label Propagation. Supervised models are trained using tree bank information with SHMM-ME: Covered TBD used tree bank tag set for the words only if they are also in the Wiktionary and All TBD uses tree bank tag sets for all words. 50, 100 and All Sent. models are trained in a supervised manner using increasing numbers of training sentences.

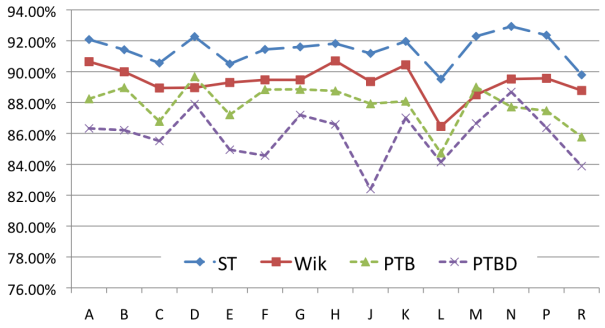

Figure 6: Model accuracy across the Brown corpus sections. ST: Stanford tagger, Wik: Wiktionary-tag-set-trained SHMM-ME, PTBD: PTB-tag-set-trained SHMM-ME, PTB: Supervised SHMM-ME. Wik outperforms PTB and PTBD overall.

5.6 Error breakdown

In Section 3.4 we discussed the accuracy of the Wiktionary tag sets and as Table 2 shows, a dictionary with better tag set quality generally (except for Greek) improves the POS tagging accuracy. In Figure 7, we group actual errors by the word type classified into the five cases discussed above: identical, superset, subset, overlap, disjoint. We also add oov – out-of-vocabulary word types. The largest source of error across languages are out-of-vocabulary (oov) word types at around 45% of the errors, followed by tag set mismatch types: subset, overlap, disjoint, which together comprise another 50% of the errors. As Wiktionary grows, these types of errors will likely diminish.

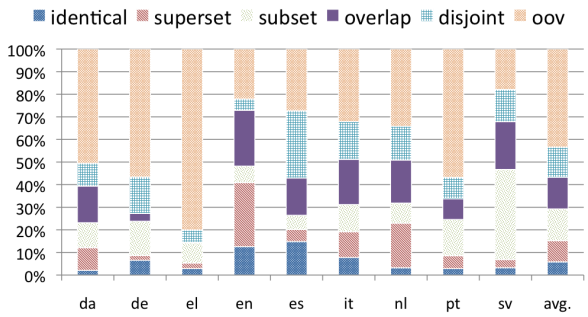

Figure 7: Tag errors broken down by the word type classified into the six classes: oov, identical, superset, subset, overlap, disjoint (see text for detail). The largest source of error across languages are out-of-vocabulary (oov) word types, followed by tag set mismatch types: subset, overlap, disjoint.

6 Conclusion

We have shown that the Wiktionary can be used to train a very simple model to achieve state-of-art weakly-supervised and out-of-domain POS taggers. The methods outlined in the paper are standard and easy to replicate, yet highly accurate and should serve as baselines for more complex propos-

als. These encouraging results show that using free, collaborative NLP resources can in fact produce results of the same level or better than using expensive annotations for many languages. Furthermore, the Wiktionary contains other possibly useful information, such as glosses and translations. It would be very interesting and perhaps necessary to incorporate this additional data in order to tackle challenges that arise across a larger number of language types, specifically non-European languages.

Acknowledgements

We would like to thank Slav Petrov, Kuzman Ganchev and André Martins for their helpful feedback in early versions of the manuscript. We would also like to thank to our anonymous reviewers for their comments and suggestions. Ben Taskar was partially supported by a Sloan Fellowship, ONR 2010 Young Investigator Award and NSF Grant 1116676.

References

A. Abeillé. 2003. *Treebanks: Building and Using Parsed Corpora*. Springer.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Proc. NAACL*, June.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

S. Buchholz and E. Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics.

S.F. Chen. 2003. Conditional and joint models for grapheme-to-phoneme conversion. In *Proc. ECSCT*.

P. Chesley, B. Vincent, L. Xu, and R.K. Srihari. 2006. Using verbs and adjectives to automatically classify blog sentiment. *Training*, 580(263):233.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proc. EACL*.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609, Portland, Oregon, USA, June. Association for Computational Linguistics.

Weiwei Ding. 2011. Weakly supervised part-of-speech tagging for chinese using label propagation. Master's thesis, University of Texas at Austin.

Jianfeng Gao and Mark Johnson. 2008. A comparison of Bayesian estimators for unsupervised hidden Markov model POS taggers. In *In Proc. EMNLP*, pages 344–352, Honolulu, Hawaii, October. ACL.

S. Goldwater and T. Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *In Proc. ACL*, volume 45, page 744.

J.V. Graça, K. Ganchev, L. Coheur, F. Pereira, and B. Taskar. 2011. Controlling complexity in part-of-speech induction. *Journal of Artificial Intelligence Research*, 41(2):527–551.

J. Graça, K. Ganchev, F. Pereira, and B. Taskar. 2009. Parameter vs. posterior sparisty in latent variable models. In *Proc. NIPS*.

A. Haghighi and D. Klein. 2006. Prototype-driven learning for sequence models. In *Proc. HTL-NAACL*. ACL.

M Johnson. 2007. Why doesn't EM find good HMM POS-taggers. In *In Proc. EMNLP-CoNLL*.

AA Krizhanovsky and F. Lin. 2009. Related terms search based on wordnet/wiktionary and its application in ontology matching. *Arxiv preprint arXiv:0907.2209*.

Michael Lamar, Yariv Maron, Mark Johnson, and Elie Bienenstock. 2010. SVD and clustering for unsupervised POS tagging. In *Proceedings of the ACL 2010 Conference: Short Papers*, pages 215–219, Uppsala, Sweden, July. Association for Computational Linguistics.

Yoong Keok Lee, Aria Haghighi, and Regina Barzilay. 2010. Simple type-level unsupervised POS tagging. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 853–861, Cambridge, MA, October. Association for Computational Linguistics.

M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

B. Merialdo. 1994. Tagging English text with a probabilistic model. *Computational linguistics*, 20(2):155–171.

C. Müller and I. Gurevych. 2009. Using wikipedia and wiktionary in domain-specific information retrieval.

Evaluating Systems for Multilingual and Multimodal Information Access, pages 219–226.

E. Navarro, F. Sajous, B. Gaume, L. Prévot, H. ShuKai, K. Tzu-Yi, P. Magistry, and H. Chu-Ren. 2009. Wiktionary and nlp: Improving synonymy networks. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 19–27. Association for Computational Linguistics.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*. Association for Computational Linguistics.

J. Nocedal and Stephen J. Wright. 1999. *Numerical optimization*. Springer.

S. Petrov, D. Das, and R. McDonald. 2011. A universal part-of-speech tagset. *Arxiv preprint ArXiv:1104.2086*.

A. Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proc. EMNLP*. ACL.

Sujith Ravi and Kevin Knight. 2009. Minimized models for unsupervised part-of-speech tagging. In *In Proc. ACL*.

H. Schütze. 1995. Distributional part-of-speech tagging. In *Proc. EACL*, pages 141–148.

Libin Shen, Giorgio Satta, and Aravind Joshi. 2007. Guided learning for bidirectional sequence classification. In *Proc. ACL*, Prague, Czech Republic, June.

N. Smith and J. Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proc. ACL*. ACL.

B. Snyder, T. Naseem, J. Eisenstein, and R. Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1041–1050. Association for Computational Linguistics.

K. Toutanova and M. Johnson. 2007. A Bayesian LDA-based model for semi-supervised part-of-speech tagging. *In Proc. NIPS*, 20.

K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *In Proc. HLT-NAACL*.