Leveraging Machine Readable Dictionaries in Discriminative Sequence Models

Ben Wellner and Marc Vilain

The MITRE Corporation
Bedford MA, USA
wellner@mitre.org

Abstract

Many natural language processing tasks make use of a *lexicon* – typically the words collected from some annotated training data along with their associated properties. We demonstrate here the utility of *corpora-independent lexicons* derived from machine readable dictionaries. Lexical information is encoded in the form of features in a Conditional Random Field tagger providing improved performance in cases where: i) limited training data is made available ii) the data is case-less and iii) the test data genre or domain is different than that of the training data. We show substantial error reductions, especially on unknown words, for the tasks of part-of-speech tagging and shallow parsing, achieving up to 20% error reduction on Penn TreeBank part-of-speech tagging and up to a 15.7% error reduction for shallow parsing using the CoNLL 2000 data. Our results here point towards a simple, but effective methodology for increasing the adaptability of text processing systems by training models with annotated data in one genre augmented with general lexical information or lexical information pertinent to the target genre (or domain).

1. Introduction

While many standard text processing tasks, such as part-of-speech tagging, shallow parsing (chunking), named entity identification and others, have reached high levels of performance on standard datasets, systems customized for such datasets often generalize poorly to different text styles, genres or domains. For example, (Tsuruoka, Tateishi et al. 2005) show that performance of a state-of-the-art part-of-speech tagger trained on the PennTree Bank WSJ corpus achieves only 85.2% accuracy on the GENIA corpus from the biomedical domain.

Some recent work has looked at this general problem of improved robustness or adaptability. While one thread of research, referred to as *transfer learning*, aims at developing new, or augmenting existing, machine learning methods (cf. (Sutton and McCallum 2005)), another approach lies in improving background resources for these systems and their ability to properly utilize such resources.

In this work, we follow the latter path by developing corpora independent lexicons derived from the Collins' English Dictionary and encoding various *lexicon features* in a Conditional Random Field for the tasks of part-of-speech tagging and shallow parsing. Our experiments point towards a methodology whereby existing part-of-speech taggers and shallow parsers can be made robust in the face of new genres or domains through the use of lexicons. This potentially ameliorates the need for expensive manual annotation in the target genre or domain.

2. Background

2.1. Lexicon Features for Tagging

Many approaches to sequence labeling tasks such as part-of-speech tagging have made use of lexicons or lexicon-based features. Initial work on Transformation-based Learning for part-of-speech tagging (Brill 1995) leveraged wordlists derived from a portion of the training data to determine default part-of-speech tags. Wordlists have been used in Hidden Markov Models for named entity tagging (Burger, Henderson et al. 2002) whereby words occurring infrequently in the training data are replaced by a category label. Recent work using Conditional Random Fields (CRFs) also makes considerable use of lexicon features (McDonald and Pereira 2005). However, no work of which we are aware has attempted to identify the contribution of lexicon features with a specific eye towards robustness across genres, domains or with degraded data quality such as case-less data.

2.2. Conditional Random Fields

Conditional Random Fields (CRFs) are conditionally-trained finite state machines that have been applied very successfully to a wide range of tasks including both POS tagging (Lafferty, McCallum et al. 2001; Cohn, Smith et al. 2005) and shallow parsing (Sha and Pereira 2003). CRFs provide the contextual advantages of sequence-based models such as HMMs (in the form of dependencies between adjacent labelings) with the discriminative power of classifiers and flexibility to include arbitrary features of the observed data (e.g. the words in a sentence). The probability of $\boldsymbol{y}$, a sequence of labels, given $\boldsymbol{x}$, a sequence of observations (e.g. words) is given by:

$$P(y \mid x) = \frac{\sum_i \sum_k \lambda_k f_k(x, y_{i-1}, y_i)}{Z_x}$$

The $f_k(x_i,y_{i-1},y_i)$ are arbitrary feature functions (with associated parameters λ_k over the input sequence observations, $\boldsymbol{x}$, and the current, y_i, and previous, y_{i-1}, label. For example, a feature can query whether the word at position i in the sequence is part of a lexicon, LEX_{NN}:

if $x_i \in LEX_{NN}$ and y_{i-1} = "DT" and y_i = "NN" then $f_k(x_i,y_{i-1},y_i) = 1$; otherwise $f_k(x_i,y_{i-1},y_i) = 0$

It is worth noting that in contrast to standard approaches with HMMs, features regarding lexicon membership do not "fire" only for rare words, but for *all* words. Thus, more robust statistics regarding the different word-list-based features can be gathered from the training data. These features are clearly highly inter-dependent with the word features themselves. For example, the lexicon feature associated with adverbs, LEX_{RB} will be very highly correlated with the word "quickly". A conditional model, such as a CRF, can accommodate such highly dependent features – something that is not possible without drastic modeling assumptions in HMMs, for example.

2.3. Alternative Training and Inference Methods for Conditional Random Fields

One drawback of CRFs is their high computational cost at training time. The time and space complexity for *inference* is quadratic in the number of labels, linear in the length of each sequence and linear in the number of features (assuming all features are cached and not recomputed). With many iterations required during training, training times can quickly become unmanageable since inference over the entire training set must be performed at each iteration. For applications such as part-of-speech tagging, this can be especially problematic as there are many states (possible labels) in the model.

In order to handle larger state spaces with large amounts of data, a number of approximations – or variations – to CRFs have been explored in the literature. Maximum Entropy Markov Models (McCallum, Freitag et al. 2000) avoid performing inference during training time by instead using the observed states in the data for computing per-state marginal probabilities. Cyclic Dependency Networks (Toutanova, Klien et al. 2003; Tsuruoka and Tsujii 2005) use the same idea with a bi-directional model that allows for a state label to be influenced both by states preceding and following the current state. Another recent approach is the use of error-correcting codes to find an optimal label sequence from a set of CRFs, each of which covers a small subset of the possible labels (Cohn, Smith et al. 2005). This can reduce the complexity in the number of states from quadratic to linear with small reductions in accuracy.

Another alternative introduced for sequence labeling in (Collins 2002) is the Averaged Perceptron which avoids the many iterations of training required by maximizing the conditional log-likelihood. Instead, parameters are adjusted with a simple update determined by the difference in the observed frequency of a feature in a sequence compared with the frequency of that feature in the maximum a posteriori (MAP) state (using the current parameters, determined using the Viterbi). Typically, just a few passes through the entire data set are required. Instead of taking the parameter values after the final pass, however, the average values of the parameters (over each epoch) are taken as the final model parameters. This averaging helps to avoid overfitting.

3. Methods

3.1. CRF Tagger

Our CRF implementation, Carafe[1], handles both Averaged Perceptron and standard (conditional) maximum likelihood training. For the task of part-of-speech tagging, we employed Averaged Perceptron training as the large state space (there are 45 possible parts-of-speech in the Penn TreeBank) made maximum likelihood training prohibitive[2]. The shallow parsing task has a smaller state space, however, so we employed full conditional log-likelihood training. Training times for part-of-speech tagging were on the order of 5-8 hours, (roughly 1 Averaged Perceptron epoch per hour), and 3-4 hours for shallow parsing using maximum likelihood learning.

Decoding using Carafe was on the order of 3000 words/sec for part-of-speech tagging and 55,000 words/sec for shallow parsing using a 3.0 GHz machine with full Viterbi decoding.

3.2. Lexicon Features

The wordlists used for lexicon features were constructed by assigning a word to the list corresponding to its part-of-speech in the CED. Words with multiple parts-of-speech were assigned to multiple wordlists. The categories include: NN, NNP, VB, VBN, VBG, JJ (adjectives), RB (adverbs) and UH (interjections).

During processing, lexicon features are triggered by looking up each word and introducing a feature corresponding to the lists that word was found in.

4. Experiments

4.1. Part-of-speech Tagging

We evaluated the use of lexicon features derived from the CED for part-of-speech tagging using the standard training, devtest and test splits over the WSJ section of the PennTreeBank (Sections 0-18 training, 19-21 devtest and 22-24 testing) using both caseful and upcase-only versions of the data. In addition, we report results of applying our models trained on WSJ to the Brown section of the Penn TreeBank. The Brown corpus contains texts from a variety of different genres considerably different than that of the WSJ.

For the part-of-speech tagging task, our CRF incorporates many of the standard features found in the literature (Ratnaparkhi 1996; Toutanova, Klien et al. 2003) including various contextual features such as local *n*-grams, word prefixes and suffixes, and tag bi-gram features. The complete set of features types is listed in Table 1.

Table 2 shows the token-level accuracy of the standard and lexicon-enhanced systems trained on sections 0-18 of WSJ and tested on both the WSJ development set (Sections 19-21) and the entire Brown corpus. Additionally, we trained and tested both the baseline and lexicon-enhanced systems using only upcase data on both the WSJ devtest and Brown test sets. Case-less data is relatively common in a variety of domains where texts have been automatically transcribed or converted from a legacy format.

Without case, the system is stripped of valuable orthographic clues and, among other things, distinguishing proper nouns and common nouns becomes considerably more difficult. We hypothesized, and our experiments

[1] Carafe is available open source, implemented in Objective Caml, at http://sourceforge.net/projects/carafe

[2] Full maximum likelihood training using L-BFGS optimization takes on the order of 7-8 days over sections 0-18 of the WSJ.

confirmed, that the lexicon-enhanced systems are especially effective in this setting based on the intuition that an unknown word not appearing in a broad-coverage lexicon of common nouns is more likely to be a proper noun.

Current tag	t_i
Current and previous tags	t_{i-1}, t_i
Word unigrams, current tag	w_i, t_i ; w_{i-1}, t_i ; w_{i-2}, t_i ; w_{i-3}, t_i ; w_{i+1}, t_i ; w_{i+2}, t_i ; w_{i+3}, t_i
Word bigrams, current tag	w_{i-1}, w_i, t_i ; w_i, w_{i+1}, t_i ; $w_{i-1}, w_{i+1} t_i$
Word bigram, current and previous tags	$w_{i-1}, w_i, t_{i-1}, t_i$
Trigrams, current state	$w_{i-2}, w_{i-1}, w_i, t_i$; $w_{i-1}, w_i, w_{i+1}, t_i$
Suffixes	(Suffixes of w_i), t_i
Prefixes	(Prefixes of w_i), t_i
Lexical Features	(w_i has a hyphen), t_i (w_i has all capital letters), t_i (w_i has an initial capital), t_i (w_i has a number), t_i
*Lexicon Feature	(w_i belongs to *LEX*$_j$), t_i

Table 1: Features used for part-of-speech tagging. The lexicon feature (*) is only used for the "Lexicon Enhanced" experimental runs.

	Standard	Lexicon-enhanced	Error reduction
WSJ-w/case	97.04%	97.18%	4.73%
WSJ-no case	96.16%	96.83%	17.45%
Brown-w/case	94.95%	95.51%	11.09%
Brown-no case	93.62%	94.19%	8.93%

Table 2: Part-of-speech tagging accuracy on the WSJ development set (Sections 19-21) and the Brown corpus.

	Number of unkn. words	Standard; unkn. word	Lexicon-enhanced; unkn. word	Error reduction
WSJ-w/case	4467	87.37%	88.94%	12.4%
WSJ-no case	3862	84.52%	88.11%	23.19%
Brown-w/case	27570	82.52%	86.18%	20.94%
Brown-no case	24576	75.68%	80.56%	20.06%

Table 3: Part-of-speech unknown word accuracy on the WSJ and Brown corpora.

We also measured the effect lexicon features had on out-of-vocabulary accuracy. This is exactly where we would expect the most gain from incorporating corpora-independent lexicons. Table 3 shows the overall accuracy on out-of-vocabulary words on the WSJ test set and the Brown corpus for both caseful and case-less versions of the data.

Finally, we measured the effect of the lexicon-enhanced features with reduced training set sizes, demonstrating the greater utility of lexicons in the face of small amounts of training data as shown in Table 4.

# of training sentences	Standard	Lexicon-enhanced	Error reduction
1000	92.37%	93.89%	19.9%
5000	95.56%	96.01%	10.1%
10000	96.17%	96.43%	6.7%
20000	96.72%	96.89%	5.2%

Table 4: WSJ devtest performance using reduced amounts of training data.

Our part-of-speech tagging performance on the WSJ test set (Sections 22-24), using the lexicon-enhanced system, was 97.21% which compares favorably with the best published results of 97.24% (Toutanova, Klien et al. 2003). Our approach however, makes a simpler first-order Markov assumption, rather than a second-order one used there.

4.2. Shallow Parsing

We performed a similar set of experiments on the task of shallow parsing (chunking) as defined in the CoNLL 2000 shared task (Tjong-Kim-Sang and Buchholz 2000).

Following standard practice, we use the BI encoding of chunk tags (Ramshaw and Marcus 1995). That is, each word is assigned a tag indicating whether it is the beginning or inside of a particular chunk type. There are 12 different chunk types in the CoNLL scheme including the "O" chunk which represents words outside of any chunk (e.g. punctuation). With both B and I type tags, this amounts to 24 states, still markedly fewer than for part-of-speech tagging.

Our shallow parsing system uses a very similar set of features to that of the part-of-speech tagger including various n-grams not only of the words but of the assigned part-of-speech tags when provided. We trained the shallow parser on the entire WSJ training set as well as just sections 15-18 as was done in the CoNLL 2000 shared task. Testing was performed on section 20 of the WSJ following CoNLL.

We intentionally *excluded* part-of-speech information as features to better measure the effectiveness of the lexicon-enhanced system and to determine how well a shallow parser could perform without part-of-speech information. A high-performing shallow parser, not requiring part-of-speech tags as input, is of great interest not least because part-of-speech tagging is slow and avoiding it allows for much faster text processing.

Training Data	Standard	Lexicon-enhanced	Error Reduction
Sections 15-18	88.13	90.00	15.75%
Sections 0-18	92.04	92.76	9.05%

Table 6: Shallow parsing F-measure on Section 20 of the WSJ without part-of-speech tags as input.

The results for shallow parsing are shown in Table 6. Results are stated in terms of F-measure – balanced precision and recall of chunk tags. The lexicon-enhanced system performs better when training on sections 15-18 as well as on the entire WSJ training set, though the error-reduction is less when training on the entire WSJ training set.

As a point of comparison, our shallow parser achieves 93.77% F-measure when using part-of-speech tag information as a source for features (with part-of-speech tags generated from the lexicon-enhanced part-of-speech tagger described above). We found this result somewhat surprising since the part-of-speech tagger was trained on the same data as the shallow parser (Sections 0-18 of WSJ). This indicates that part-of-speech tagging has some inherent utility: accurate assignment of syntactic categories to individual words appears to improve the generalization capability of a shallow parser. Nevertheless, the lexicon-enhanced shallow parser without using part-of-speech tags performs very respectably at a much higher rate of throughput.

5. Conclusion

Part-of-speech tagging and shallow parsing are critical components in many language processing tasks. While performance on standard data sets is now very high, systems that generalize well to new domains, genres or reduced data quality remain elusive. In this paper we have demonstrated a simple approach towards significantly improving generalization performance on such tasks with the incorporation of lexicon features derived from machine readable dictionaries – reducing error across genres and within.

We demonstrated this improvement on part-of-speech tagging by showing reduced error rates with the use of lexicon features on the WSJ devtest set. More extensive error reductions were achieved on the Brown corpus and with case-less data. The benefit of lexicon features was also apparent in the face of limited training data. The story was similar for shallow parsing, where we also obtained significant reductions in the error rate using lexicon features.

These results suggest that the use of corpora-independent lexicon features can serve as useful features in discriminative sequence models. Further exploration applying the approach advocated in this paper to specific domains such as biomedicine and other areas with more specific, but documented nomenclatures is a promising area for future work.

Acknowledgements

We thank John Burger for helpful comments along with the anonymous reviewers. This work was supported under MITRE Sponsored Research.

References

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.

Burger, J., J. C. Henderson, et al. (2002). Statistical Named Entity Tagger Adaptation. Conference on Natural Language Learning (CoNLL-2002), Taipai, Taiwan.

Cohn, T., A. Smith, et al. (2005). Scaling Conditional Random Fields Using Error-Correcting Codes. Proceedings of the Association of Computational Linguistics, 2005., Ann Arbor, MI, USA.

Collins, M. (2002). Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. Proceedings of EMNLP, Philadelphia PA, USA.

Lafferty, J. D., A. McCallum, et al. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequences. ICML 01: Proceedings of the Eithteenth International Conference on Machine Learning, San Francisco CA, USA, Morgan Kaufmann.

McCallum, A., D. Freitag, et al. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. Proc. 17th International Conf. on Machine Learning, San Francisco CA USA.

McDonald, R. and F. Pereira (2005). "Identifying Gene and Protein Mentions in Text Using Conditional Random Fields." BMC Bioinformatics 6((Suppl 1):S13).

Ramshaw, L. A. and M. P. Marcus (1995). Text Chunking Using Transformation-Based Learning. Proceedings of the Third ACL Workshop on Very Large Corpora, Cambridge, MA, USA.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-of-Speech Tagging. Proceedings of EMNLP.

Sha, F. and F. Pereira (2003). Shallow parsing with conditional random fields. Proceedings of HLT-NAACL 2003.

Sutton, C. and A. McCallum (2005). Composition of Conditional Random Fields for Transfer Learning. Proceedings of Human Language Technologies/Emperical Methods in Natural Language Processing Conference (HLT/EMNLP 2005), Vancouver, B.C., Canada.

Tjong-Kim-Sang, E. F. and S. Buchholz (2000). Introduction to the CoNLL-2000 Shared Task: Chunking. Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal.

Toutanova, K., D. Klien, et al. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. Proceedings of HLT-NAACL.

Tsuruoka, Y., Y. Tateishi, et al. (2005). Developing a Robust Part-of-Speech Tagger for Biomedical Text. Advances in Informatics - 10th Panhellenic Conference on Informatics, LNCS.

Tsuruoka, Y. and J. i. Tsujii (2005). Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data. HLT/EMNLP.