Evaluating and improving morpho-syntactic classification over multiple corpora using pre-trained, "off-the-shelf", parts-of-speech tagging tools

Kevin Glass and Shaun Bangay

Department of Computer Science
Rhodes University
Grahamstown
[k.glass;s.bangay]@ru.ac.za

Abstract

This paper evaluates six commonly available parts-of-speech tagging tools over corpora other than those upon which they were originally trained. In particular this investigation measures the performance of the selected tools over varying styles and genres of text without retraining, under the assumption that domain specific training data is not always available. An investigation is performed to determine whether improved results can be achieved by combining the set of tagging tools into ensembles that use voting schemes to determine the best tag for each word. It is found that while accuracy drops due to non-domain specific training, and tag-mapping between corpora, accuracy remains very high, with the support vector machine-based tagger, and the decision tree-based tagger performing best over different corpora. It is also found that an ensemble containing a support vector machine-based tagger, a probabilistic tagger, a decision-tree based tagger and a rule-based tagger produces the largest increase in accuracy and the largest reduction in error across different corpora, using the Precision-Recall voting scheme.

1. Introduction

1.1. Problem Statement

We investigate the performance of "off-the-shelf" parts-of-speech taggers over a variety of different corpora. In particular we investigate the previously un-investigated problems of:

- how well commonly available parts-of-speech taggers perform over test-data other than that upon which they are trained, to access the expected performance without retraining the tool for a specific domain;
- whether combining commonly available parts-of-speech taggers into vote-based ensembles leads to a reduction in error, and if so, the determination of the set of tools that produce the best results, as well as the determination of the type of voting scheme that most reduces error.

1.2. Background

The process described in this paper is a component of a Text-to-Scene conversion system, where information extracted from fiction books is used to populate three-dimensional virtual worlds with objects and movements. Descriptions of dramatic scenes occur frequently in fiction books, describing scene contents and layout, and the movements and actions of the entities in the scene. Identifying the parts-of-speech associated with each word in the text is an important part of accomplishing this task. For example a word that corresponds to an Object that appears in a scene must be used as a *noun* in the sentence.

Many "off-the-shelf" tools are available for the common natural language processing task of morpho-syntactic classification, otherwise known as word-class disambiguation or parts-of-speech tagging. In general these tools are concerned with assigning the correct class (noun, verb, adjective) or *tag* to each word in the input text. The majority of these tools employ automated classification techniques to achieve this goal, and require training over a corpus of natural language known to be accurately tagged with parts-of-speech.

The standard training corpus for this task is the Wall Street Journal section of the Penn Treebank [1], and this is also the standard benchmark for determining the accuracy of each tool. However, while such evaluations provide best-case indications of the accuracy of these taggers, they do not indicate the results that should be expected using these tools over various styles and genres of natural language text. Practical use of parts-of-speech tagging tools should not assume that domain specific training is feasible. As such we evaluate the performance of a suite of available parts-of-speech taggers over different styles and genres of natural language text, with the restriction that no retraining is performed and using only the trained models distributed with the tools.

It is expected that parts-of-speech tagging tools may not perform at reported rates of accuracy when used over different styles and genres of natural language, specifically without the ability to retrain each tool. It is expected that this error may be countered through the use of ensembles of parts-of-speech tagging tools, where the word-class for each token is determined by allowing each tagger in the ensemble to vote for their preferred candidate. The combination of tagging tools that produces the largest reduction in error needs to be determined, as well as the method for performing voting.

This paper investigates the success of different parts-of-speech tagging techniques over not only the Wall Street Journal, but other corpora such as the Lancaster/Oslo-Bergen (LOB) corpus [2] and the Brown corpus [1, 3], both containing English from a variety of sources ranging from technical writings and newspaper articles to excerpts from fiction books. In addition, the LOB corpus makes use exclusively of British English, while the Brown corpus consists only of American English.

1.3. Overview

This paper is structured as follows: existing evaluation and combination research is described in Section 2, while a description of the techniques used to evaluate and combine parts-of-

speech taggers over multiple corpora is given in Section 3. Section 4 presents the results of the validation and combination experiments, followed by conclusions in Section 5.

2. Related Work

Evaluation of parts-of-speech tagging tools over different corpora is described in existing work [4, 5]. In these experiments each tagger is retrained specifically for each corpus, eliminating experimental error which may results from language and tag-set differences between taggers and corpora. As such these studies do not provide a view of expected accuracy where retraining is not possible.

Initial work by the same authors [6] tests available parts-of-speech taggers (and ensembles thereof) over the limited SUSANNE [7] corpus, finding a definitive decrease in individual tagger accuracy, and limited success with simple voting schemes with tagger ensembles. However, van Halteren *et al.* [8, 4] and Marquez *et al.* [9] show that the correct use of voting schemes with ensembles may indeed result in accuracies higher than any one tagger in the ensemble.

The primary problem with using multiple corpora and multiple "off-the-shelf" parts-of-speech tagging tools is the use of differing (and often conflicting) tag-sets. Table 1, presents the details of the corpora used for testing in this paper. As indicated, the tagging scheme of the Penn Treebank contains a total of 48 different tag categories (36 excluding punctuation), while the LOB corpus employs a set of 153 tags, and the SUSANNE 353. It cannot be guaranteed that a parts-of-speech tagger uses the same set of tags of each corpus, which makes direct evaluation of a tagger over different corpora impossible.

	Tokens	Tokens (excl. punct.)	Full Tag-set	Tag-set (excl. punct.)
SUSANNE	50 325	42 889	353	341
WSJ	1 288 623	1 114 957	48	36
Brown	1 170 775	1 015 425	48	36
LOB	1 157 220	997 906	153	141

Table 1: Size, content and tag-sets for four test corpora.

Solutions have been proposed for *mapping* tag-sets onto one another, either manually or automatically [10, 11, 12] but this is difficult since in many cases tagging schemes are not mappable. For instance, mappings of type $1 : n$ and $n : m$ occur, where in the former case, a single tag type in one scheme maps to n tag types in another tag-set [10].

This research differs from existing work in that a set of taggers is evaluated over multiple corpora without retraining for any specific tag-set. Instead an independent tag-set is created to which the other tag-sets are mapped. The set of tagging tools used in this study have not previously been evaluated in this manner, and have not been previously combined into ensembles.

3. Evaluation and combination over multiple corpora

For the experiments in this paper, six different taggers are chosen, each using a different class of classification technique[1]. This is based on the premise that although each tagger uses the same contextual information regarding the current word to define its tag, each one makes use of it in a different manner. Table 2 lists the taggers used for this research, the underlying classification mechanism and the reported accuracy of each. Note that the reported accuracies of each tagger cannot be fairly compared in this manner since many are evaluated using different corpora, and different tag-sets. Information regarding classification techniques is summarised by Glass and Bangay [6], or in the accompanying publication for each tagger.

Name	Type	Reported Accuracy
QTag [13]	Probabilistic	98.39%
TreeTagger [14]	Decision Tree	96.36%
Brill Tagger [15]	Rule-based	97.20%
Stanford Tagger [16]	Maximum Entropy	97.24%
SVMTool [17]	Support Vector Machine	97.20%
POS Tagger [18]	Bidirectional Perceptron	97.33%

Table 2: Freely available parts-of-speech taggers, and the accuracy reported for each.

To overcome the obstacle of different tag-sets used by the corpora presented in Table 1, a coarse independent tag-set is created to which the other tag-sets are mapped. Creating this map is non-trivial, and in many cases tags cannot be accurately mapped onto the coarse set accurately. Avoiding these cases is impossible, and where such flawed mappings are easily identified, any sentence in the corpus containing such a tag is removed from the test corpus (we found that this problem only occurs with the SUSANNE corpus, which explains why it is reported as having 50 325 tokens in Table 1, as against its actual 130 000). In some cases detection of $1 : n$ or $n : m$ mappings are less obvious, especially given the lack of expert knowledge of each tagging scheme. These cases are ignored, and assumed to be a part of experimental error under the assumption that majority of a corpus falls under well-mapped tags. The coarse tag-set developed for this research is a simplified version of the tag-set applied in the Penn Treebank, and is presented in Table 3.

All taggers in Table 2 make use of the Penn tag-set, with the exception of QTag, which uses a modified version of the LOB tag-set. As a result, tags assigned by these tools are also mapped to the coarse tag-set presented in Table 3.

Two different experiments are performed using the six taggers over each corpus presented in Table 1:

- **Validation:** to validate the reported accuracies of the discussed parts-of-speech taggers on a variety of corpora, without retraining.
- **Combination:** to determine if a higher accuracy can be achieved by combining this novel collection of parts-of-speech taggers into a voting system.

3.1. Validation of Tagger Accuracy

The aim of this experiment is to validate the accuracy of each of the above mentioned parts-of-speech taggers over the various presented corpora. Taggers are not retrained, and use the models with which they are distributed.

[1] At time of writing, the Association of Computational Linguistics list the POS Tagger, Stanford Tagger and SVMTool as the top three parts-of-speech taggers available. Source: `http://aclweb.org/aclwiki/` [accessed on 26 September 2007].

	Tag	Description
1.	CC	Co-ordinating Conjunction
2.	CD	Cardinal Number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
6.	IN	Preposition/subordinating conjunction
7.	JJ	Adjective
8.	MD	Modal
9.	NN	Noun
10.	NNP	Proper Noun
11.	PRP	Pronoun
12.	RB	Adverb
13.	RP	Particle
14.	TO	To
15.	UH	Interjection
16.	VB	Verb
17.	VBD	Verb, past tense
18.	VBG	Verb, gerund or present participle
19.	VBN	Verb, past participle
20.	VBZ	Verb, third person singular present
21.	WDT	Wh-determiner
22.	WP	Wh-pronoun
23.	WRB	Wh-adverb
24.	@COPY@	*Punctuation*

Table 3: Coarse tag-set, adapted from the Penn [1] tag-set.

A *gold standard* is created by mapping the original tags in each corpus to the coarse tag-set. Initially each corpus is stripped of tags, the result of which is passed through each parts-of-speech tagger. The output of each tagger is also mapped to the coarse tag-set for validation. Each tag in the automatically tagged corpus is compared to the corresponding tag in the gold standard, and *accuracy* refers to the percentage of correct tags with regard to the total number of tokens in the corpus. Note that tagging of punctuation is not included in any validation.

3.2. Combination of parts-of-speech taggers

The purpose of this experiment is to determine if more accurate tagging can be achieved by combining the different parts-of-speech taggers into voting systems. In particular, the experiment must determine which type of voting scheme results in the largest reduction in error, as well as which combination of parts-of-speech taggers produces the highest accuracy.

A number of voting schemes are investigated, in which the tag suggested for a word by each tagger in the ensemble is weighted in a different manner:

1. **Simple Vote:** the tag which is selected by majority of the taggers is chosen. In the case of a tie, a random tag from those suggested by the tied parties is chosen.
2. **Weighted Vote:** each tagger contributes a specified weighting in support of a particular tag. The tag with the highest cumulative score is chosen. The following weighting schemes are evaluated:

   (a) *TotAccuracy:* each tagger has a weighting equivalent to its overall accuracy, determined as a result of the validation experiments described in Section 3.1.

   (b) *TagPrecision:* according to the weighting scheme defined by van Halteren *et al.* [4], a tag specific weighting scheme may be employed by making use of the *precision* of the specific tagger, with regards to any particular tag. Precision for any tag χ is the percentage of of tokens tagged χ by the tagger that are also tagged thus in the gold standard:

   $$precision = \frac{number\ \chi\ tags\ correct}{number\ \chi\ tags\ assigned\ by\ tagger}$$

   Precision values for each tag assigned by a specific tagger are calculated as a byproduct of the validation experiments described in Section 3.1.

   (c) *Precision-Recall:* according to the weighting scheme defined by van Halteren *et al.* [4], a tag-specific weighting scheme may be employed that not only takes into account how successful a particular tagger is at tagging a certain tag of type χ, but also the error that the other taggers in the ensemble experience when assigned a tag of type χ. Error can be derived from a tagger's *recall* rate, where recall for any tag χ is the percentage of tokens tagged χ in the gold standard that are also tagged χ by the tagger:

   $$recall = \frac{number\ \chi\ tags\ correct}{number\ \chi\ tags\ in\ gold\ standard}$$

   Recall values for each tag assigned by a specific tagger are calculated as a byproduct of the validation experiments described in Section 3.1. Error is calculated as $(1 - recall)$, and measures how often a tagger fails to recognise a specific tag. The weighting assigned by tagger τ for tag χ in this scheme is therefore calculated as follows (where S is the set of taggers in the ensemble):

   $$weight_{\chi}^{\tau} = precision_{\chi}^{\tau} + \sum_{\forall \lambda \in S/\tau} error_{\chi}^{\lambda}$$

   where the set S/τ is the set of taggers, excluding tagger τ.

3. **Ranked Vote:** each of the taggers is given a rank (between 1 and 6) based on the performance in Section 3.1, with the best scoring tagger assigned a rank of six. Tag scores are calculated by adding the ranks of the taggers which voted for each specific tag. The tag which achieves the highest score is chosen.

As indicated by van Halteren *et al.* [4] *second level learners* (machine learning techniques that learn models of optimal tag selection from an ensemble of parts-of-speech taggers) may be used on top of an ensemble of taggers for improved results, but these techniques do not perform well when there is a lack of training data. As such second level learners are not incorporated into this experiment.

The validation experiment in Section 3.1 results in tagged corpora for each parts-of-speech tagger. These files are used as input into the combination experiments. Initially a plain-text version of each corpus is created. This file is traversed in parallel with the various automatically tagged corpora, token by token. The tag assigned to a specific token by each tagger is used

in the voting schemes described above in order to determine the most suitable tag. It is also expected that certain ensembles of taggers will perform better than others, and so every possible combination of the six taggers is evaluated.

Accuracy is used as a metric for evaluating ensembles of taggers (see Section 3.1). Additionally an improvement in error is also calculated as follows (based on the metric defined by van Halteren *et al.* [4]):

$$err = \frac{correct_{ensemble} - correct_{best}}{total - correct_{best}} * 100$$

This metric indicates the reduction in error (as a percentage) achieved using an ensemble from the error produced using the best performing tagger in the ensemble (calculated in Section 3.1). For instance, if an ensemble contains two taggers where the best accuracy of the two is 99.0%, then assuming the ensemble results in an accuracy of 99.5%, the percentage error reduction is 50%.

4. Results

4.1. Validation of Tagger Accuracy

Table 4 presents the accuracy results for the different parts-of-speech taggers. Note that the accuracies do not correspond to those reported in the literature for each tagger, and this is explained by the following points:

- **Different test data:** majority of the taggers were trained and tested over the WSJ section of the Penn Treebank, which represents only one fifth of this corpus. Results are expected to decrease as a result of larger test-beds, as well as differences in style and genre of language, and most importantly type of language (American versus British English).
- **Different tag-sets:** it is expected that a reduced tag-set such as the coarse tag-set would make the tagging task simpler, and in the case of the SVM tagger, a higher accuracy than reported results, since the coarse tag-set is a direct derivative of the Penn Tag-set, used by the SVM tagger. However, mapping from other tag-sets such as LOB and SUSANNE is likely to introduce error, which explains why all taggers produce lower accuracy scores over these corpora.

It is unclear which of the above points contribute to the majority of decrease in accuracy. However, the figures in Table 4 reflect very high accuracies in spite of these expected errors. The reduced accuracies over the LOB corpus are primarily attributed to the differences in lexicon between American and British English. Overall, the SVM tagger performs particularly well over all corpora, correctly tagging over 90% in all cases, and over 98% over the Wall Street Journal. As expected, an increase in accuracy results here because of the reduction in the size of the tag-set. The TreeTagger also performs very well, with only slightly reduced results.

It is worth noting that all taggers are trained on the WSJ, but not all produce equivalent or higher tagging accuracy over this corpus. This is attributed to the fact that the taggers are tested only on 20% of the corpus, using the remainder as training data. The results in Table 4 however reflect each taggers performance over the entire corpus.

Table 5 presents the top three contributors of error for each tagger over the LOB and Brown corpus[2]. For example, 7.71% of the total error encountered by QTag over the LOB corpus is caused by the incorrect assignment of CD instead of IN. It is evident from the table that each tagger has difficulties with certain tags across the two copora. For instance, in many cases NNP and NN tags are confused. Other major causes of confusion are NN and JJ, as well as VBD and VBN. These errors may be a result of the tag-set mapping to the coarse set, or a result of the initial training of the taggers. It is worth noting that in the case of the POS tagger, the top three errors are all caused by the same incorrectly assigned tags. This may be due to a mapping error, but is unlikely since the same mapping is used for this tagger as is used for the Stanford and SVM taggers (all three taggers use the Penn tag-set). If there were a mapping error, these three errors would be very pronounced in all three taggers.

This experiment confirms that the parts-of-speech taggers perform with a high level of accuracy over multiple types and styles of text. While accuracy is reduced in all cases, the pre-trained taggers are still capable of producing accuracies of over 90%. It remains to be determined whether ensembles of taggers may be produced to improve overall accuracy.

[2]The other two corpora are omitted due to space restrictions.

	Report	WSJ	LOB	Brown	SUS
QTag	98.39	70.40	72.87	72.51	76.077
Tree	96.36	96.94	91.67	94.45	**91.15**
Brill	97.20	93.10	88.67	92.55	88.45
Stanford	97.24	91.53	80.21	89.89	85.92
SVM	97.20	**98.17**	**92.17**	**95.10**	90.19
POS	97.33	83.36	82.59	84.74	83.02

Table 4: Accuracy results of various taggers over the different corpora

	LOB		Brown	
QTag	IN/CD	7.71%	NNP/NN	9.21%
	PRP/NN	7.67%	IN/CD	7.54%
	NNP/NN	6.59%	PRP/NN	5.99%
Tree	IN/TO	13.56%	NNP/NN	8.12%
	VB/VBG	7.25%	NN/JJ	4.84%
	NNP/JJ	3.95%	NN/NNP	4.82%
Brill	IN/TO	10.00%	IN/DT	8.92%
	IN/DT	6.85%	VBD/VBN	7.67%
	RP/IN	5.90%	VB/NN	6.57%
Stanford	NN/@COPY@	9.88%	VBD/VBN	14.02%
	IN/@COPY@	6.39%	VB/NN	8.86%
	VBD/VBN	6.21%	JJ/RB	5.44%
SVM	IN/TO	14.43%	NN/NNP	8.34%
	VB/VBG	7.49%	NN/JJ	7.66%
	NN/JJ	4.58%	JJ/NN	4.64%
POS	NNP/NN	21.14%	NNP/NN	27.02%
	VBD/VBN	7.50%	VBD/VBN	9.05%
	VB/NN	6.67%	VB/NN	6.48%

Table 5: Indication of the top three largest contributors to error for each tagger over the LOB and Brown corpus

4.2. Combination of parts-of-speech taggers

Table 6 presents the top results achieved by different ensembles over the four test corpora. In particular, the table lists the

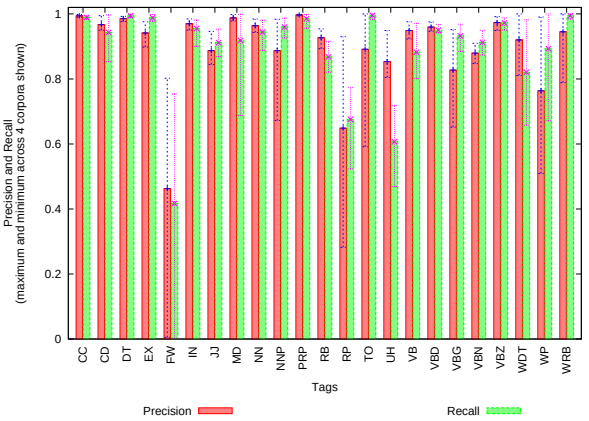

Figure 1: Precision and recall value for each tag using the SRPT ensemble over the LOB corpus. Maximum and minimum values over all 4 corpora are indicated using error-bars.

ensemble that achieves the highest accuracy (Acc.), as well as the ensemble that achieves the highest reduction in error (Err.). Note that for each corpus a decrease in error is achieved using an ensemble, and in all cases a higher accuracy is achieved from using an ensemble than from using any one of its individual component taggers. Table 6 also shows that a reduction in error of over 32% can be achieved using ensembles. In the case of the Brown corpus, the same ensemble achieves the highest accuracy as well as the highest improvement in error. Thus for illustrative purposes, the `srpt` ensemble is also shown.

It is evident from Table 6 that not all ensembles produce an improvement. In many cases the accuracy of an ensemble is reduced from the highest individual accuracy of its component taggers. This is always the case for the simple voting scheme. However, the Precision-Recall voting scheme suggested by van Halteren *et al.* [4] consistently produces the most reduction in error and an increase in accuracy over all the component taggers, over all corpora.

Table 6 also provides evidence in support of the consistently best ensemble. In particular the ensemble containing the SVM tagger, the Rule-based tagger, the Probabilistic tagger and the Tree tagger (`srpt`) provides the best improvements, and highest accuracies over the LOB and SUSANNE corpus. The same ensemble, excluding the Probabilistic tagger (`srt`) produces the highest increase in score and error reduction over the Brown corpus. However, including QTag into the ensemble results in a comparably high accuracy and increase in error reduction.

Assuming future processing will use the best performing ensemble of taggers, it is useful to determine expected levels of success regarding the assignment of individual tags. Figure 1 presents the mean precision and recall value for each tag in the coarse tag-set obtained using the `srpt` ensemble with the Precision-Recall voting scheme, over the four different corpora. The maximum and minimum values are also indicated using error-bars. The figure indicates high levels of precision and recall over majority of the tags. However, the ensemble produces poor results over foreign words (FW), particles (RP) and interjections (UH), where error bars reach 0% in some cases, indicating that none of these tags were correctly assigned over one of the corpora. Hence future processes of the Text-to-Scene conversion system should not rely on the accurate identification of these classes of words.

				Simple		TotAccuracy		TagPrecision		Ranked		Precision-Recall	
Corpus	Maximum:	Ensemble	Best	Acc.	Err.	Acc.	Err.	Acc.	Err.	Acc.	Err.	Acc.	Err.
WSJ	Accuracy	`sp`	98.17	85.00	-717.86	98.17	0	98.07	-5.35	98.17	0	98.17	**0.44**
	Error Reduction	`px`	83.36	76.78	-39.52	83.36	0	84.63	**7.64**	83.36	0	85.53	**13.06**
LOB	Accuracy	`srpt`	92.17	91.53	-8.18	92.45	**3.67**	92.73	**7.21**	92.45	**3.61**	92.75	**7.51**
	Error Reduction	`ep`	80.21	76.29	-19.83	80.21	0	85.55	**26.96**	80.21	0	86.63	**32.43**
SUSANNE	Accuracy	`srpt`	91.15	90.98	-1.95	91.09	-0.74	91.23	**0.82**	91.03	-1.34	91.68	**5.96**
	Error Reduction	`px`	83.02	81.39	-9.58	83.02	0	85.72	**15.89**	83.02	0	85.42	**14.13**
Brown	Accuracy and Error Reduction	`srt`	95.1	95.76	**13.48**	95.75	**13.22**	95.85	**15.23**	95.75	**13.22**	95.82	**14.62**
	(second-best)	`srpt`	95.1	94.62	-9.87	95.75	**13.3**	95.72	**12.69**	95.75	**13.31**	95.75	**13.23**

Table 6: Summary of highest combination scores and error improvements over four different corpora. Ensembles are indicated using groupings of the following letters: {s}VM tagger, {p}robabilistic tagger, {r}ule based tagger, {t}ree tagger, Max-{e}nt tagger, POS Tagger {x}. Error reductions are indicated in bold.

This experiment demonstrates that error resulting from single parts-of-speech taggers can be significantly reduced using the correct ensemble of parts-of-speech taggers, and the correct choice of voting scheme. In particular, the `srpt` ensemble is recommended for parts-of-speech tagging tasks, using the Precision-Recall voting scheme defined by van Halteren *et al.* [4].

5. Conclusion

It is concluded that parts-of-speech taggers perform well without any further training beyond what is shipped with the software. This section shows that, in particular, the SVM tagger performs consistently well over different types and styles of text. The result of this research is the choice of the `srpt` ensemble of parts-of-speech taggers, consisting of the pre-trained SVMTool, Brill Tagger, QTag, and Tree-Tagger software. In addition, the Precision-Recall voting scheme defined by van Halteren *et al.* [4] is chosen as the scheme most likely to reduce error encountered by individual taggers in the ensemble.

Contributions of this work include extension of the work of van Halteren *et al.* [4] by testing ensembles of taggers over different corpora, without retraining, and confirms that ensembles of "off-the-shelf" taggers are also capable of producing more accurately tagged text. In addition, the work presented here extends the previous research presented by Glass and Bangay [6] by evaluating novel set of taggers and ensembles over different corpora, including the Wall Street Journal and Brown sections of the Penn Treebank, as well as the Lancaster-Oslo/Bergen corpus.

6. References

[1] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1994.

[2] Stig Johansson, Eric Atwell, Roger Garside, and Geoffrey Leech, *The Tagged LOB Corpus: Users' manual*, ICAME, The Norwegian Computing Centre for the Humanities, Bergen University, Norway, 1986, http://www.comp.lancs.ac.uk/ucrel/local/lob/ [accessed on 10 September 2007].

[3] W. N. Francis and H. Kucera, *Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers.*, Brown University, Providence, Rhode Island, U.S.A., 2 edition, July 1979, http://icame.uib.no/brown/bcm.html [accessed on 10 September 2007].

[4] Hans van Halteren, Walter Daelemans, and Jakub Zavrel, "Improving accuracy in word class tagging through the combination of machine learning systems," *Comput. Linguist.*, vol. 27, no. 2, pp. 199–229, June 2001.

[5] Simone Teufel, Helmut Schmid, Ulrich Ileid, and Anne Schiller, "Study of the relation between tagsets and taggers," Tech. Rep., Expert Advisory Group on Language Engineering Standards, University of Stuttgart, Germany, May 1996, EAGLES Document EAG-CLWG-TAGS/V.

[6] Kevin Glass and Shaun Bangay, "Evaluating parts-of-speech taggers for use in a text-to-scene conversion system," in *Proceedings of SAICSIT '05*, Judith Bishop and Derrick Kourie, Eds., White River, South Africa, September 2005, pp. 20–28.

[7] Geoffrey Sampson, "The SUSANNE analytic scheme," http://www.grsampson.net/RSue.html [accessed on 10 September 2007].

[8] Hans van Halteren, Jakub Zavrel, and Walter Daelemans, "Improving data driven wordclass tagging by system combination," in *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, Christian Boitet and Pete Whitelock, Eds., San Francisco, California, 1998, pp. 491–497, Morgan Kaufmann Publishers.

[9] Luis Márquez, Horacio RodrÃguez, Josep Carmona, and Josep Montolio, "Improving pos tagging using machine-learning techniques.," in *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999, pp. 53–62.

[10] Simone Teufel, "A support tool for tagset mapping," in *Proceedings of the Workshop SIGDAT (EACL95)*, Dublin, Ireland, March 1995, European Chapter of the Association for Computational Linguistics.

[11] Eric Atwell, John Hughes, and Clive Souter, "AMALGAM: Automatic mapping among lexico-grammatical annotation models," in *The Balancing Act: Combining Symbolic and Statistical Approaches to Language - Proceedings of the ACL Workshop*, J. Klavans, Ed. 1994, pp. 21–28, Association for Computational Linguistics.

[12] Hervé Déjean, "How to evaluate and compare tagsets? a proposal," in *Proceedings of LREC2000*, Athens, Greece, 2000.

[13] Dan Tufis and Oliver Mason, "Tagging romanian texts: a case study for qtag, a language independent probabilistic tagger," in *Proceedings First LREC*, Granada, Spain, 1998.

[14] Helmut Schmid, "Probabilistic part-of-speech tagging using decision trees," Tech. Rep., IMS, Univ. of Stuttgart, 1994.

[15] Eric Brill, "Some advances in transformation-based part of speech tagging," in *AAAI '94: Proceedings of the twelfth national conference on Artificial intelligence (vol. 1)*, Menlo Park, CA, USA, 1994, pp. 722–727, American Association for Artificial Intelligence.

[16] Kristina Toutanova and Christopher Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, Hong Kong, 2000.

[17] Jesús Giménez and Lluís Marquez, "Fast and accurate part-of-speech tagging: The SVM approach revisited.," in *RANLP*, 2003, pp. 153–163.

[18] Libin Shen, Giorgio Satta, and Aravind Joshi, "Guided learning for bidirectional sequence classification," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, Prague, Czech Republic, June 2007, Association of Computational Linguistics, pp. 760–767, Association of Computational Linguistics.