Training and evaluation of POS taggers on the French MULTITAG corpus

A. Allauzen, H. Bonneau-Maynard

LIMSI/CNRS; Univ Paris-Sud, Orsay, F-91405
{allauzen,maynard}@limsi.fr

Abstract

The explicit introduction of morphosyntactic information into statistical machine translation approaches is receiving an important focus of attention. The current freely available Part of Speech (POS) taggers for the French language are based on a limited tagset which does not account for some flectional particularities. Moreover, there is a lack of a unified framework of training and evaluation for these kind of linguistic resources. Therefore in this paper, three standard POS taggers (Treetagger, Brill's tagger and the standard HMM POS tagger) are trained and evaluated in the same conditions on the French MULTITAG corpus. This POS-tagged corpus provides a tagset richer than the usual ones, including gender and number distinctions, for example. Experimental results show significant differences of performance between the taggers. According to the tagging accuracy estimated with a tagset of 300 items, taggers may be ranked as follows: Treetagger (95.7%), Brill's tagger (94.6%), HMM tagger (93.4%). Examples of translation outputs illustrate how considering gender and number distinctions in the POS tagset can be relevant.

1. Introduction

The most widely used French Part of Speech (POS) tagger is the French version of the Treetagger (Schmid, 1994) which is freely available on the web[1]. A version of the Brill's tagger (Brill, 1994), trained on the GRACE (Paroubek et al., 1998) corpus is also frequently cited[2]. Both taggers are trained with small tagsets (around 50 tags) which do not include number or gender distinction. Moreover, no confident comparative evaluation of this taggers has been performed yet since they are based on two different tagsets and trained in different conditions.

Developing a corpus-based POS tagger relies on two main resources. On one hand, the tagger itself is a set of machine learning algorithms. On the second hand, the training data consists in a (semi-)manually annotated corpus which defines the tagset : the set of POS classes that the tagger aims to assign. The French Part-Of-Speech tagging evaluation GRACE project was performed on a text of 20k words extracted from the French newspaper *Le Monde*. This text was manually tagged using 50 different tags. The compared systems (Symbolic or corpus based) were trained and developed with their own linguistic resources[3].

The present work has been motivated by the development of a new French-English statistical machine translation (SMT) system which includes morphosyntactic knowledge. This work requires a French POS tagger trained with a tagset larger than 50 tags, with a richer representation of the typical French inflections. For example, gender and number distinctions can be useful to disambiguate the translation of English ambiguous words which yield to different forms in French. Adjectives and participle past are typical examples of this phenomenon. Therefore, this paper presents the development and the evaluation of statistical POS taggers for French on a same corpus using the same tagset: the French MULTITAG (Paroubek, 2000) corpus, which is a by-product of the GRACE project.

This large corpus (more than 840k words) includes a very large tagset (more than 1500 tags). As one of our goal is to provide a comparative evaluation, a part of the corpus is excluded from the training data to provide an unseen test set. For this experiment, three state-of-the-art statistical taggers are trained: the Brill's tagger (Brill, 1994), Treetagger (Schmid, 1994) and a standard Hidden Markov Model (HMM) tagger (Charniak et al., 1993).

This paper is organized as follows. Next section addresses the feasible integration of POS information in SMT systems. The section 3. provides an overview of the three tested taggers. The content of the MULTITAG corpus is then described, along with the normalization process. The last section presents and discusses the experimental results and provides examples of the possible impact of POS knowledge on SMT outputs.

2. POS information for statistical machine translation

Recent works in statistical machine translation (SMT) show how phrase-based modeling significantly outperforms the historical word-based modeling. Using phrases, i.e. sequences of words, as translation units allows the system to preserve local word order constraints and to improve the consistency of phrases during the translation process. As opposed to word-based models, phrase-based models provide some sort of context information and implicitly capture syntactic and semantic relations.

However the output of a SMT system is often difficult to understand by humans requiring re-ordering words and recovering its syntactic structure. It is well-known that syntactic structures vary greatly across languages. French or Spanish, for example, can be considered as highly inflectional languages, whereas inflection plays only a marginal role in English. Therefore, explicit introduction of syntactic structure of the language in statistical models becomes a promising focus of attention.

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[2] The tagger can be downloaded from http://research.microsoft.com/˜brill/ and the French version is available on the web site of INALF: http://www.inalf.cnrs.fr/scripts/mep.exe?HTML˜mep_winbrill.txt

[3] Grace evaluation home page: http://www.limsi.fr/TLP/grace/

In a recent work (Bonneau-Maynard et al., 2007), the introduction of morphosyntactic information into a phrase based SMT model was explored, by enriching words with their morphosyntactic categories. In this case, it seems likely that the morphosyntactic information of each word is useful to encode linguistic characteristics, resulting in a sort of word disambiguation by considering its morphosyntactic category. Encouraging results have been obtained for translation from English to Spanish, on the TC-STAR task (public European Parliament Plenary Sessions translation). Further experiments are underway to evaluate a tighter integration of morphosyntactic information in SMT such as the use of factored model (Koehn and Hoang, 2007).

Morphosyntactic information has also been successfully introduced in SMT to perform word reordering, as proposed in (Popovic and Ney, 2006) or in (Crego et al., 2006) for the language pair Spanish-English. Therefore, a preprocessing reordering step is done before training and translation in both source and target language sequences.

3. POS taggers

Three different POS taggers are used in the reported experiments. Our selection is threefold motivated. These taggers use an statistical approach. They yield to state of the art results. And last, they are freely available and distributed with all the necessary training tools.

A sentence of n words can be considered as a sequence of random variables $\boldsymbol{W} = w_1...w_n$. Statistical POS tagging aims to associate to $\boldsymbol{W}$, a sequence of random variables $\boldsymbol{T} = t_1...t_n$, where t_i represents the POS tag assigned to the word w_i. In the Bayesian approach, the goal is to find $\boldsymbol{T}^*$ that maximizes the posterior probability:

$$\boldsymbol{T}^* = \operatorname*{argmax}_{T} p(\boldsymbol{T}|\boldsymbol{W}) = \operatorname*{argmax}_{T} p(\boldsymbol{W}|\boldsymbol{T})p(\boldsymbol{T}) \qquad (1)$$

Two questions arise to develop a statistical POS tagger: the question of learning or how to estimate the terms $p(\boldsymbol{W}|\boldsymbol{T})$ and $p(\boldsymbol{T})$, and the question of decoding or how to find the best sequence $\boldsymbol{T}^*$ given a new word sequence. The learning phase is based on a training corpora which is a set of couples $(\mathbf{W},\mathbf{T})$. The training data are known to be always too small and sparse hence the need of assumptions about the statistical dependencies among the random variables involved in equation 1. The POS taggers that are used in this experiment can be distinguished by these assumptions.

3.1. Classical HMM tagger

The classical HMM tagger is fully described in (Charniak et al., 1993) and makes the following Markovian assumptions:

$$p(\boldsymbol{W}|\boldsymbol{T}) = \prod_{i=1}^{n} p(w_i|t_i) \qquad (2)$$

$$p(\boldsymbol{T}) = \prod_{i=1}^{n} p(t_i|t_{i-1}) \qquad (3)$$

The first assumption means that the occurrence of a word only depends on its associated tag (observation probabilities), and the second that a tag can be completely predicted knowing its previous tag or the bigram transition probabilities. Despite these simplifications, smoothing methods must be used to deal with data sparseness as proposed in (Charniak et al., 1993). Therefore the training process aims to estimate the transition and observation probabilities. To answer the decoding question, the best tag sequence is assigned using the standard Viterbi algorithm. This algorithm is for example described for the POS tagging task in (Manning and Schütze, 1999).

3.2. Treetagger

The Treetagger assumes trigram transition probabilities :

$$p(\boldsymbol{T}) = \prod_{i=1}^{n} p(t_i|t_{i-1}, t_{i-2})$$

To deal with data sparseness, the trigram probabilities are estimated by growing a decision tree.

3.3. Brill's tagger

The Brill's tagger (Brill, 1994) starts with a more simple assumption: each word is first labeled with its most probable POS tag based on the training corpus. This first and raw POS tagging is then corrected with sequencing transformation rules. These rules are learned from the training corpus and encode various and complex inter-dependencies between words and tags. A specific rule set is also dedicated to the prediction of POS tags for unknown words (unseen during the training step). This last kind of rules are not used for the following experiments.

4. Corpus

For English, two well-known POS tagged corpus are usually used to train POS taggers: the Brown Corpus (Francis and Kucera, 1982) and the Penn Treebank (Marcus et al., 1994). For French, there are no such widely used linguistic resources.

4.1. Corpus description

The GRACE French Part-Of-Speech tagging evaluation project (Paroubek et al., 1998) was carried on a 20k word corpus. Text data were extracted from articles of the French newspaper *Le Monde*. These texts were manually tagged using 50 different tags. Even if a version of the Brill's Tagger trained by INALF on this corpus is already freely available, it appears that the tagset is not large enough for the investigated application. For example there is no gender or number distinction. The problem seems to be similar for the corpus on which the French version of Treetagger was trained.

The MULTITAG (Paroubek, 2000) corpus is a by-product of the GRACE project. This 1 million word corpus has been produced by a Rover combination of the data produced by the systems which participated to the GRACE evaluation. The Rover combination consists in a voting strategy to select the correct annotation among the hypotheses provided by the systems (Fiscus, 1997). A manual correction has been performed only on annotations on which systems did not converge (no majority vote). The MULTITAG corpus size - 840k words (30k sentences) - is very promising for statistical training.

Another interest of this corpus is the exceptionally large tagset. Since the objective of the GRACE project was to evaluate many different systems, the final tagset had to ensure the compatibility between all participants and their specific tokenization. The resulting unified tagset consists of 1500 different tags. The MULTITAG tagset includes a dozen of lexical categories (`Noun`, `verb`, `adjective`...). For each category, several subcategories with their corresponding values are defined. For example for the `Noun` category three attributes or subcategories are defined: `type` with the corresponding values `common`, `proper` and `cardinal`, `Gender`, with the corresponding values `feminine` and `masculine`, and `Number` with the corresponding values `singular` and `plural`.

4.2. Corpus normalization

The text normalization process aims to define what is considered to be a word. Although normalization may result in a reduction of information, it typically reduces ambiguity and redundancy, and this step cannot be helped for data sparsity compensation. In this work, the usual processing steps are performed such as the ambiguous punctuation marks (such as hyphens and apostrophes) or the sentence initial capitalization.

Moreover, in the MULTITAG corpus, frequent word sequences and named entities are split with specific tags. For example, the French sequence of words *"au cours de"* for the English word *"during"* appears in the corpus as *au Sp/1.3 cours Sp/2.3 de Sp/3.3*, which means that this sequence contains three words and has the syntactic role of a preposition. To be coherent with machine translation normalization, this word sequence has to be converted in a single compound word. On the other side, sequences like the French named entity *"Président de la République"* (*"President of the Republic"*) have to be split. Due to normalization issues, some sentences were discarded.

The final corpus contains about 600k words for 27k sentences with a final tagset of 300 items. A reduced version of the tagset has been considered to assess the impact of data sparseness. The criterion to simplify the tagset was to keep the categories, the gender and number distinctions and to discard some information about sub-categories such as the mood or tense for verbs, type or degree for adjectives. The resulting reduced tagset contains 130 different labels.

5. Evaluation

To provide a test set, 2500 sentences are randomly sampled from the final corpus. The rest of the data is used to train the POS taggers. The taggers are evaluated in terms of tagging accuracy using the held out test data.

5.1. Quantitative performances

The results reported in Table 1 show that the performances (95.7% tagging accuracy for the best system) are quite similar to the usually reported results for English data. For example on the English Penn Treebank corpus, the tagging accuracy is about 97.2% with the Brill's tagger, and 96.7% with the standard HMM. To explain the loss in performance between French and English for both of these taggers, one might consider that in English the evaluation was performed under the closed vocabulary assumption and with a smaller tagset. Thus one can observe that, for French, the Treetagger outperforms the Brill's tagger with a significant absolute difference of 1.1% in tagging accuracy, and the HMM tagger with a difference of 2.3%.

Tagger	Tagging accuracy
HMM	93.4%
Brill	94.6%
Treetagger	95.7%

Table 1: Tagging accuracy obtained by the three taggers on the 2500 sentences test set, using the tagset of 300 items

To assess the impact of the tagset on the tagging accuracy, a similar experiment was carried out using the same data but using a reduced set of 130 tags. While the overall tagging accuracy increases of 1%, the reduction of the tagset did not modify the ranking of the POS taggers.

The same trend is observed when comparing the precision and recall for the gender subcategorization. The use of Treetagger results in a precision of 96.7% and a recall of 96.2% compared with the precision of 94.9% using Brill's tagger and its associated recall of 94.9%. These performances are close to the overall tagging accuracy. Whereas the precision measures are similar for the number distinction, the recall measures are significantly lower and about 80% for both taggers.

5.2. Qualitative analysis of errors

A manual analysis of errors shows recurrent confusions such as:

- the decision concerning the annotation of the verbs *"être"* (to be) and *"avoir"* (to have) between auxiliary or verb tags.
- confusion between tags for adjectives and participle past (which can be used as adjective in some context),
- tagging of numbers, which can be partially solved with a specific normalization,
- the important ambiguity for several words like *"que"*, or *"des"*.

One can observe Brill's tagger systematically attributes a Proper Name tag to words beginning with a capital. This last problem could be corrected with the Brill's tagger by learning or adding new morphological rules to guess the tags for unknown words.

5.3. Translation examples

Different SMT systems are currently under development using POS tags with factored translation models. Although quantitative evaluation is not yet available, the examples of the figure 1 show how gender and number can be helpful for translation. For both examples, the source sentence in English is first given. Then translations coming from three SMT systems are given. The first translation system corresponds to a standard phrase-based SMT system, using

English:	this needs to be said to all _those who are asserting_ the opposite
Baseline translation:	cela doit être dit à tous _ceux qui *sont affirmant*_ le contraire
Translation with 50 tags	cela doit être dit à tous _ceux qui *sont affirmant*_ le contraire
Translation with 130 tags:	cela doit être dit à tous _ceux qui *affirment*_ le contraire
English:	the problem is that , _if you set a date_ , there is a danger
Baseline translation:	le problème est que , _si *vous fixer* une date_ , il existe un risque
Translation with 50 tags	le problème est que ,_si *vous fixer* une date_ , il existe un risque ...
Translation with 130 tags:	le problème est que , _si *nous fixons* une date_ , il existe un risque
English:	_whatever the economic progress_ made , whatever the social progress in Tunisia
Baseline translation:	_*quelles* que soient les progrès économiques_ réalisés , quel que soit le progrès social en Tunisie
Translation with 50 tags	_*quelles* que soient les progrès économiques_ réalisés , quel que soit le progrès social en Tunisie
Translation with 130 tags:	_*quel* que soit le progrès économique_ , quel que soit le progrès social en Tunisie

Figure 1: Comparative translations using the baseline phrase-basesd SMT system and two systems enhanced with POS information. The second translation system is enhanced with units composed of words enriched with POS tags coming from the standard version of the Treetagger (i.e. with a 50 tag tagset) whereas the third system uses POS tags obtained with the Brill's tagger trained on the MULTITAG corpus (i.e. with a tagset of 130 items including gender and number distinctions).

words as units. The second translation system is enhanced with units composed of words enriched with POS tags coming from the standard version of the Treetagger (i.e. with a 50 tag tagset) whereas the third system uses POS tags obtained with the Brill's tagger trained on the MULTITAG corpus (i.e. with a tagset of 130 items including gender and number distinctions described in subsection 4.2.).

In the first example, the baseline system outputs _"ceux qui sont affirmant"_ which is not syntactically correct. The same translation is also produced by the translation system based on the small tagset. A better translation is obtained with the third system that may be attributed to the number constraint linking the subject _"ceux qui"_ - which is plural - to the verb form _"affirment"_ - which is the correct plural form. In the second example, the same phenomenon is observed: the incorrect form _"si vous fixer une date"_ produced by the baseline and the first translation system, does not appear in the last translation where the verb form _"fixons"_ agrees in number (first plural person) with the subject _"nous"_.

Examples corresponding to gender errors are less frequent. In the third example, a gender error can be observed in the baseline and the first translation system hypothesis, whereas the gender agreement is correct with the last system between the noun _"progrès"_ which is masculin and the pronoun _"quel"_.

6. Conclusion

Three POS taggers for French have been trained and evaluated on the same large corpus MULTITAG. Their performances were compared using 2500 sentences as test set. Results show that the performance of the taggers can be ranked as follow: the best tagger is the Treetagger, followed by the Brill's tagger, and both of them outperform the standard HMM tagger. Nevertheless, the conception of the Brill's tagger allows the user to easily improve or adapt an already-trained tagger to a new domain or a new type of corpus. Adaptation can be performed by simply adding well suited rules to include knowledge about out-of-vocabulary words or particularities of the corpus such as tokenization or named entities. This kind of flexibility is not possible with the Treetagger.

The next step will be to evaluate the usability of each tagger in a phrase based SMT experiment using factored models (Koehn and Hoang, 2007). Preliminary examples show that, in the case of translating from English to French, the use of a tagset including gender and number is efficient in correcting some translation errors.

7. Acknowledgment

This work has been partially financed by OSEO under the Quaero program. The authors wish to thanks Noemie Colin for her contribution to this work and Patrick Paroubek for his help.

8. References

H. Bonneau-Maynard, A. Allauzen, D. Déchelotte, and H. Schwenk. 2007. Combining morphosyntactic enriched representation with n-best reranking in statistical translation. In _proc. Syntax and Structure in Statistical Translation (SSST), NAACL-HLT 2007 / AMTA Workshop_, April.

E. Brill. 1994. Some advances in rule based part-of-speech tagging. In AAAI, editor, _Proceedings of the Twelfth National Conference on Artificial Intelligence_, pages 722–727, Seattle, WA.

Eugene Charniak, Curtis Hendrickson, Neil Jacobson, and Mike Perkowitz. 1993. Equations for part-of-speech tagging. In _National Conference on Artificial Intelligence_, pages 784–789.

Josep M. Crego, Adrià de Gispert, Patrik Lambert, Marta R. Costa-jussà, Maxim Khalilov, Rafael Banchs, José B. Mariño, and José A. R. Fonollosa. 2006. N-gram-based smt system enhanced with reordering patterns. In _Proceedings on the Workshop on Statistical Machine Translation_, pages 162–165, New York City, June. Association for Computational Linguistics.

J. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (rover).

W. Nelson Francis and Henry Kucera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin Company.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876.

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

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Patrick Paroubek, Josette Lecomte, Gilles Adda, Joseph Mariani, and Martin Rajman. 1998. The grace french part-of-speech tagging evaluation task. In *First International Conference on Language Resources and Evaluation (LREC)*, pages 433–441, May.

Patrick Paroubek. 2000. Language resources as by-product of evaluation: the multitag example. In *Second International Conference on Language Resources and Evaluation (LREC) 2000*, pages 151–154.

Maja Popovic and Hermann Ney. 2006. Pos-based word reorderings for statistical machine translation. In *5th International Conference on Language Resources and Evaluation (LREC)*, pages 1278–1283, May.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, September.