DEVELOPMENT OF A MULTILINGUAL PARALLEL CORPUS AND A PART-OF-SPEECH TAGGER FOR AFRIKAANS

Julia Trushkina
Centre for Text Technology,
North-West University,
2531 Potchefstroom, South Africa
20215770@puk.ac.za

Abstract

This paper describes design and creation of a multilingual parallel corpus for South African languages. One of the applications of the corpus, namely, the induction of a part-of-speech tagger for Afrikaans from the data, is presented in the paper. Development of the Afrikaans part-of-speech tagger is based on a modified method for induction of linguistic tools from parallel corpora originally proposed by Yarowsky and Ngai (2001).

Keywords: Natural Language Processing, Parallel corpora, induction of linguistic tools, South African languages, Afrikaans, Part-of-Speech tagging.

1. Introduction

Multilingual annotated corpora, such as the Multext (Ide and Veronis, 1994) and the Multext-East (Dimitrova et al., 1998) corpora, are among the most valuable resources in current natural language processing. They underlie statistical research in multilingual tasks, such as machine translation, multilingual lexicography and word sense disambiguation, and can also be used in projects on monolingual studies.

For multilingual communities, such as the community of South Africa with eleven official languages, creation of a multilingual corpus has a special significance. It provides a basis for the development of multilingual language applications that can be used to facilitate or even avoid labor- and time-consuming processes of manual handling of multilingual information.

Additionally, such a corpus enables empowerment of minority languages of multilingual communities. With the use of a parallel corpus and the meth-

Please use the following format when citing this chapter:

Trushkina, J., 2006, in IFIP International Federation for Information Processing, Volume 228, Intelligent Information Processing III, eds. Z. Shi, Shimohara K., Feng D., (Boston: Springer), pp. 453–462.

ods which allow the transfer of linguistic annotations across languages, new resources and tools can be created for the minority languages.

The goal of the research project presented in this paper is the development of a multilingual corpus and basic tools and resources for South African languages. The current paper describes creation of such multilingual corpus and a development of a part-of-speech (POS) tagger for Afrikaans, one of the most prominent languages in South Africa. Although a member of the Indo-European family, Afrikaans is a language with very few resources. Several collections of unannotated Afrikaans texts exist, but the only corpus with incorporated linguistic information currently available for Afrikaans is a small corpus of approximately 20 000 tokens annotated with POS analyses (Pilon, 2006).

For the development of a POS tagger for Afrikaans, we apply a modified method of induction of linguistic tools from parallel data originally described in (Yarowsky and Ngai, 2001). project can be easily employed for additional development of tools for other South African languages.

2. Potchefstroom Bible Corpus

Different sources of multilingual texts have been discussed in the literature. They include, among others, collections of law documents, such as the Canadian Hansard and the collection of European Parliamentary documents, translations of novels and other fiction, and multilingual versions of web pages (Resnik, 1999).

In the current project, the text of the Bible has been chosen as the basis for the multilingual corpus. The motivation of this choice is twofold. First, the Bible is available in many languages and is often accessible in electronic format, even for such rare languages as Maori and Swahili[1]. This makes the future expansion of corpus to other languages possible. The second reason for selecting the Bible as the content of the corpus is the close correspondence of the Bible translations in different languages.

At present, the corpus comprises the Bibles in five languages: Afrikaans, isiZulu, isiXhosa, English and Dutch. The first four languages are the most widely spoken languages in South Africa. An additional reason for the inclusion of the English data into the corpus is the high variety of freely available resources for English which can be used in annotation transfer. Dutch, the only language of the corpus which is not an official language of South Africa, has been included in the corpus since it is the closest relative of Afrikaans, which can make the transfer of linguistic analysis to Afrikaans more accurate.

The following Afrikaans, English and Dutch translations of the Bible have been chosen: the 1983 version of the Afrikaans translation, the World English Bible, and the Dutch Statenvertaling Bible. The choice of these versions has

been motivated by two considerations: the modern language of the texts and the availability of the full text in machine-readable format. The size of the corpus ranges between 820 000 and 840 00 tokens for different languages.

The Afrikaans, English and Dutch parts of the corpus have been aligned on sentence and word level with freely available tools.[2] The Vanilla aligner (Danielsson and Ridings, 1997) has been used for sentence alignment, whereas word alignment has been performed with the GIZA software (Och and Ney, 2003).

Sentence Alignment

With the use of Vanilla aligner, optimal sentence alignments have been found for each pair of the Indo-European languages of the corpus. The results of the automatic alignment have been checked and corrected manually. Next, bilingual alignments have been combined into trilingual alignments. The principle of maximal span has been used for the combination: the span of the resulting trilingual aligned chunks of text corresponds to the span of the *"maximal"* pair of aligned sentences. Thus, for example, if Afrikaans-Dutch alignment is 2:1 (two Afrikaans sentences to one Dutch sentence) and corresponding Dutch-English alignment is 1:1, the resulting trilingual alignment is 2:1:1.

Word Alignment

For the word alignment of the corpus data, the GIZA software has been used. The software represents one of the open-source tools developed at the EGYPT project (Och et al., 1999) for machine translation. GIZA aligner relies on a statistical method based on co-occurrence of words of different languages in aligned sentences (Model 3 of the IBM statistical machine translation formalism (Brown et al., 1990).

GIZA produces only many-to-one alignments, i.e. any word of a source language can be aligned maximally with one word in a target language. The opposite situation, in which several words of a source language are linked to a single word in a target language, is possible. Since both many-to-one and one-to-many alignments occur in natural language, we have produced two alignments for each pair of the Indo-European languages of the corpus, assuming different translation directions in the experiments. The word alignment incorporated in the Potchefstroom Bible corpus is a combination of the six alignments obtained in this way. The combination has been performed in several steps.

First, the intersection of alignments for each language pair has been assumed to be a *"safe"*, or *"reliable"* alignment. Second, semi-automatic heuristics have been implemented to increase the number of reliable alignments. By semi-automatic nature of heuristics we mean the following: candidates for re-

liable alignments are proposed by a heuristic automatically, but a confirmation of a human is required for the inclusion of the candidate into the list of reliable alignments.

The following heuristics have been used:

- *Transitivity* heuristic:

  If reliable alignments exist between word W_a of language A and W_b of language B, as well as between word W_b and word W_c of language C, then a candidate reliable alignment between W_a and W_c is proposed, given that a link $W_a - W_c$ has been established in one of the six alignment experiments.

  $$W_a - W_b,\ W_b - W_c \longrightarrow W_a - W_c$$

- *Inter-span* heuristic:

  Let ${W^a}_{n-1}$, ${W^a}_n$ and ${W^a}_{n+1}$ be a sequence of words in language A, and ${W^b}_{k-1}$, ${W^b}_k$ and ${W^b}_{k+1}$ be a sequence of words in language B. If reliable alignments exist between ${W^a}_{n-1}$ and ${W^b}_{k-1}$, as well as between ${W^a}_{n+1}$ and ${W^b}_{k+1}$, then a candidate reliable alignment between ${W^a}_n$ and ${W^b}_k$ is proposed, given that GIZA established an alignment ${W^a}_n - {W^b}_k$ in one of the six experiments.

  $${W^a}_{n-1} - {W^b}_{k-1},\ {W^a}_{n+1} - {W^b}_{k+1} \longrightarrow {W^a}_n - {W^b}_k$$

  The heuristic has been very helpful in alignment of determiners. However, human inspection of the proposed links is necessary, since in many other cases the heuristic over-applies.

- *Correction* heuristic:

  A list of common alignment errors has been compiled for the three language pairs. The most common systematic errors have been corrected manually.

  For example, the Dutch version of the Bible includes a word *"En"* in the beginning of many sentences. The Afrikaans and the English parts of the Bible more often that not do not have a corresponding conjunction in the beginning of their sentences. In such cases, the statistical module of GIZA incorrectly and systematically aligns the word *"En"* with determiners *"Die"* (in Afrikaans sentences) and *"The"* (in English sentences), because they often co-occur in the sentence pairs with *"En"*. This error is easy to identify and to correct.

The share of reliable alignments compiled in the way described above is estimated to be 57.3% for the Afrikaans-Dutch language pair and 52.38% for the Afrikaans-English language pair. A manual inspection of a small portion

of reliable alignments randomly chosen from the data demonstrated that the English-Afrikaans alignments are correct in 98.54% of cases, Dutch-Afrikaans alignments – in 98.11% of cases, and English-Dutch alignments – in 97.04% of cases.

Table 1 demonstrates an example of word-aligned data from the corpus. The first three lines represent aligned corpus sentences in Afrikaans, Dutch and English. A 6-column table under the sentences indicates alignment links for each word of the sentences.

Table 1. An example of word-aligned data from the Potchefstroom Bible corpus.

GEN 1:1 In die begin het God die hemel en die aarde geskep .
GEN 1:1 In den beginne schiep God den hemel en de aarde .
GEN 1:1 In the beginning God created the heavens and the earth

0	GEN	0	GEN	0	GEN
1	1:1	1	1:1	1	1:1
2	In	2	In	2	In
3	die	3	den	3	the
4	begin	4	beginne	4	beginning
5	het	5	schiep	6	created
6	God	6	God	5	God
7	die	7	den	7	the
8	hemel	8	hemel	8	heavens
9	en	9	en	9	and
10	die	10	de	10	the
11	aarde	11	aarde	11	earth
12	geskep	5	schiep	6	created
13	.	12	.		

3. Corpus Annotation

Analysis of the English and the Dutch Parts of the Corpus

Analysis of the English part of the Potchefstroom Bible corpus has been performed with the Charniak's parser (Charniak, 2000) – an EM parser trained on the Penn Treebank corpus (Marcus et al., 1993). The choice of the parser has been motivated by its high performance: at present, the results reported for the parser performance are the highest results for English – 90.1%. Additionally, the annotation scheme of the Penn Treebank is the most cited and widely used scheme currently employed by computational linguists working on English. The parser performs full syntactic analysis together with POS tagging. It utilizes a POS tagset of 46 tags. The syntactic analysis is based on the annotation scheme of the Penn Treebank.

The Dutch part of the corpus has been analyzed with the Alpino parser (Bouma et al., 2001) developed for Dutch at the University of Groningen. The Alpino parser provides a full syntactic analysis of Dutch together with POS annotation. It is the best parser of Dutch currently available. The results reported in the literature by the parser developers reach an accuracy of 81.3% (Bouma et al., 2001). The syntactic analysis is based on the annotation scheme of the Alpino corpus of Dutch.

4. Induction of Linguistic Analyses for Afrikaans

The annotation of the Afrikaans part of the corpus and the induction of a POS tagger for Afrikaans is based on the method proposed by Yarowsky and Ngai in (Yarowsky and Ngai, 2001).

The Method of Yarowsky and Ngai (2001)

The original model provides a high-quality annotation of a resource-poor language given a bilingual parallel corpus aligned on word level with annotation of one language part of the corpus. The method is based on an observation that linguistic analyses of translations of the same sentence in different languages often coincide.

Due to the differences in language structures and due to the often imperfect word alignments, the annotation resulting from a direct projection of analyses is of low quality. Yarowsky and Ngai (2001) report a performance of 69% for the direct projection of POS tags from English to French. The authors propose a method for robust learning from noisy POS projections by (a) down-weighting or excluding poorly aligned sentences from consideration, (b) using a bigram model for learning, (c) training the lexical prior and tag sequence models separately using generalization techniques. (Yarowsky and Ngai, 2001) report an accuracy of 97% for French using the proposed model.

Modifications to the Original Method

We follow the main principles of the described model: at first, the part-of-speech tags are projected from the English data onto the Afrikaans tokens, and then an n-gram language model is trained on the POS tag projections.

However, we modified the original model in the following ways:

1 The Afrikaans language model is trained only on reliable alignments, excluding unsafe alignments completely.

   This modification is motivated by the low quality of the automatic word alignment in our experiments.

2 To compensate for the resulting data sparseness, not only reliably aligned sentences are taken into account, as proposed in (Yarowsky and Ngai, 2001), but all safe alignments identified by the heuristics described in Section 2.2. Such safe alignments may include subsequences of sentences and even separate words.

3 A trigram model is used instead of the originally proposed bigram model.

This modification is introduced based on the generally higher performance of trigram models. Indeed, our experiments with a trigram and a bigram model have shown that the results are 1% lower for the bigram model.

4 The Afrikaans language model uses the full Penn Treebank set of 46 POS tags, unlike the originally described model which employs reduced tagsets of 14 and 9 core tags (representing main parts of speech, excluding punctuation).

5 No aggressive re-estimation of lexical probabilities in line with the original experiments is performed.

Re-estimation of lexical probabilities has been advocated in (Yarowsky and Ngai, 2001) based on the low POS ambiguity of the data used in their experiments. However, a larger tagset leads to a higher POS ambiguity of tokens, which makes the aggressive re-estimation of lexical probabilities unfavourable.

The Trigram'n'Tags (TnT) tagger, an HMM trigram tagger developed and implemented by (Brants, 2000) has been used in our tagging experiments. The TnT tagger has been trained on the corpus of reliable projections of English POS tags onto Afrikaans data. Such training corpus has a rather different structure from the structure expected by TnT for training. First, the corpus is only partially annotated, since unreliable tag projections are not included. Second, a small part of the corpus is assigned multiple tags. These multiple tags are a result of one-to-many projections, such as projections produced in case of aligning a single Afrikaans token with an English phrase.

Since the TnT tagger has not been designed to train on partially annotated data with multiple tags, the Afrikaans language model provided to TnT has been created externally: the lexicon and the n-gram statistics files have been compiled in the way described below.

All tokens with reliable alignments have been used for the creation of the TnT lexicon file. For each token, a list of POS tags associated with the token in the corpus has been produced, together with the frequencies of the token and a tag/token pair.

If an Afrikaans word has been aligned with more than one English word, tags of each English translation are included in the lexical entry of the Afrikaans

token. However, the entered frequency of such tags is reduced and represents a corresponding share of $1/n$, where n is a number of English words corresponding to the Afrikaans token.

In the creation of an n-gram statistics file, all sequences of reliably aligned text of corresponding length have been used. For example, each sequence of three words reliably aligned in the corpus has contributed to the compilation of trigrams statistics. For obtaining the statistics on unigrams, each Afrikaans word with a reliable alignment has been used.

Tagging Experiments

The TnT tagger provided with the language model compiled in the described way has been used for tagging the Afrikaans part of the corpus. The performance of the tagger has been evaluated against a manually annotated portion of the corpus. The size of the test set is 36 400 tokens. The evaluation demonstrated an accuracy of 83.98%.

When compared to the performance of the original tagger described in (Yarowksy and Ngai, 2001), the tagger induced from the Potchefstroom Bible corpus achieves a much lower accuracy. The main reason for this is a higher granularity of the tagset used in our experiments: 46 tags versus 9 tags in the original experiments.

An error analysis has demonstrated that the main sources of errors are confusion of verbal tags (32.31%), wrong tags for punctuation marks (18.06%), and mistakes that involve tag TO assigned in Penn Treebank to word *"to"* (15.28%). Mistakes in tagging of punctuation marks occur because punctuation often differs in English and Afrikaans. Table 2 presents the statistics on the occurrence of punctuation marks in the English and Afrikaans parts of the corpus. It shows a clear discrepancy in the usage of commas, full stops and semicolons. Such discrepancy leads to the projection of incorrect English tags onto Afrikaans punctuation marks.

Table 2. Statistics of the use of different punctuation marks in the Afrikaans and English parts of the corpus.

Punctuation mark	*English*	*Afrikaans*
period (.)	8 695	37 386
comma (,)	70 475	43 920
colon (:)	35 696	39 714
semicolon (;)	87 69	2 509

Errors in the use of verbal tags and the tag TO are due to the language differences of Afrikaans and English. The verbal system of Afrikaans is sig-

nificantly simpler than that of English and therefore a set of nine verbal tags that distinguish between form, tense, number and person does not make sense for Afrikaans verbs and leads to a decrease in tagging performance. Quite similarly, the use of a single tag for all translations of the English word *"to"* obviously leads to tagging errors, since it results in assigning the same analysis to a diverse group of words.

To account for these phenomena, we have performed a second experiment with a modified tagset. In the modified tagset, a single tag for all punctuation marks except for parentheses and quotes has been introduced. Verbal tags have been restricted to tags `VB` for present tense verbs and `VBD` for past participles and past tense verbs. Tag `TO` has been collapsed with the tag for prepositions (`IN`). The resulting tagset contains 33 tags. These modifications to the tagset have lead to a significant improvement of the tagging performance and resulted in an accuracy of 92.45%.

Discussion and Future Work

The proposed model for the induction of a POS tagger from parallel data represents a modified version of the original algorithm described in (Yarowsky and Ngai, 2001). The model performs training on parts of aligned sentences, including small sections of text of one or more words which the heuristics described in Section 2.2 identified as reliably linked to their counterparts in the other language.

The induced POS tagger produces analyses of high granularity. Its performance has been compared to the performance of the only existing POS tagger for Afrikaans (Pilon, 2006) – a TnT tagger trained on the small corpus of manually annotated 20 000 tokens. Both taggers have been evaluated on the same test set.

The comparison of the two Afrikaans POS taggers demonstrated that the tagger induced from the Potchefstroom Bible corpus outperforms the tagger described in (Pilon, 2006) by 10%. However, the difference in the results is influenced by the difference in tagsets employed by the two taggers. The tagset of the smaller Afrikaans corpus comprises 119 tags.

Two main directions of research on the induction of linguistic tools for Afrikaans are intended for future. The first concerns expansion of the current model to trilingual data, including the Dutch part of the corpus into experiments. The second area for future research concerns induction of other tools from the corpus data, including a noun phrase bracketer, a chunker, a named entity recognizer and a parser.

5. Conclusion

The paper described the development of a multilingual parallel corpus for South African languages, together with the experiments on the induction of a POS tagger for Afrikaans from this parallel corpus. The induction experiments have demonstrated promising results: the new POS tagger for Afrikaans outperforms a tagger trained on a small corpus of manually annotated Afrikaans corpus.

The project on the development of the corpus continues. Further development includes expansion of the corpus to other South African languages, deeper annotation of the Afrikaans part of the corpus, and alignment and linguistic analysis of the isiXhosa ans the isiZulu parts of the corpus.

Notes

1. See, for example, the Bible database website at http://www.bibledatabase.net/, which in April 2006 contained 51 versions of Bible translations in 30 languages.
2. Additional alignment of the isiZulu and the isiXhosa parts of the corpus is planned for immediate future.

References

G. Bouma, G. van Noord and R. Malouf. *Alpino: Wide-coverage Computational Analysis of Dutch*. Computational Linguistics in The Netherlands. 2001.

T. Brants. *TnT–A Statistical Part-of-Speech Tagger*. Proceedings of ANLP-2000. Seattle, 2000.

P. F. Brown, J. Cocke, S. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer and P. S. Roossin. *A Statistical Approach to Machine Translation*. Computational Linguistics 16(2):79–85, 1990.

E. Charniak. *A Maximum-Entropy-Inspired Parser*. Proceedings of ANLP/NAACL'2000. Seattle, 2000.

P. Danielsson and D. Ridings. *Practical presentation of a vanilla aligner*. Sprakbanken, Institutionen for svenska spraket, Goteborgs universitet, 1997.

L. Dimitrova, T. Erjavec, N. Ide, H.-J. Kaalep, V. Petkevic and D. Tufis. *Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages*. Proceedings of COLING'98. Montreal, 1998.

N. Ide and J. Veronis. *Multext (multilingual tools and corpora)*. Proceedings of COLING'94, p. 90–96. Kyoto, 1994.

M. Marcus, B. Santorini and M. A. Marcinkiewicz. *Building a Large Annotated Corpus of English: The Penn Treebank*. Computational Linguistics 19(2): 313–330, 1993.

F. J. Och, C. Tillmann and H. Ney. *Improved alignment models for statistical machine translation*. Proceedins of the EMNLP/WVLC Conference. 1999.

F. J. Och and H. Ney. *A Systematic Comparison of Various Statistical Alignment Models*. Computational Linguistics 29(1):19–51, 2003.

S. Pilon. *Automatic part-of-speech tagging of Afrikaans*. MA thesis, North-West University, 2006.

F. Resnik. *Mining the Web for Bilingual Text*. Proceedings of ACL'99. Maryland, 1999.

D. Yarowsky and G. Ngai. *Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora*. Proceedings of NAACL 2001. Pittsburgh, 2001.