Achieving an Almost Correct PoS-Tagged Corpus

Pavel Květoň[1] and Karel Oliva[2]

[1] Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics, Charles University,
Malostransk nm. 25, CZ - 118 00 Praha 1 - Mal Strana, Czech Republic
kveton@ufal.ms.mff.cuni.cz
[2] Austrian Research Institute for Artificial Intelligence (FAI),
Schottengasse 3, A-1010 Wien, Austria
karel@oefai.at

Abstract. After some theoretical discussion on the issue of representativity of a corpus, this paper presents a simple yet very efficient technique serving for (semi-)automatic detection of those positions in a part-of-speech tagged corpus where an error is to be suspected. The approach is based on the idea of learning and application of "invalid bigrams", i.e. on the search for pairs of adjacent tags which constitute an incorrect configuration in a text of a particular language (in English, e.g., the bigram ARTICLE - VERB). Further, the paper describes the generalization of the "invalid bigrams" into "extended invalid bigrams of length n", for any natural n, which provides a powerful tool for error detection in a corpus. The approach is illustrated by English, German and Czech examples.

1 Introduction

The quality of corpus annotation is certainly among the pressing problems in current corpus linguistics. This quality, however, is a many-faceted problem in itself, comprising both issues of a rather theoretical nature and also quite practical matters. In order to reflect this division (albeit only roughly and only within the area of part-of-speech tagging of written texts), this paper will have two layers.

In the first layer, we shall touch upon the predominantly theoretical problem of (grammatical) representativity. In the second layer, the applications of the theoretical ideas developed in the first layer will be discussed, in particular, we shall

- present some ideas concerning a method for achieving high-quality PoS-tagging
- demonstrate the practical results which were achieved when this method was applied on the NEGRA corpus ([1],[2],[6],[7]) of German.

2 Issues of Representativity of a Corpus

The notion of (grammatical) representativity of a PoS-tagged corpus of a language can be understood in at least the following two ways:

- representativity wrt. the presence of a phenomenon (broadly defined) of this language;
- representativity wrt. the relative frequency of occurrences of a (broadly defined) phenomenon in this language.

From this it is easy to conclude that representativity (in this understanding) is dependent on what is taken to be a phenomenon. (i.e. dependent on what is to be represented — not a surprising conclusion in fact In order to expose the idea of representativity and its importance in some detail, let us consider the example of representativity as needed for a training corpus for a bigram-based statistical tagger[3]. In this case, the phenomena[4] whose presence and relative frequency are at stake are:

- bigrams, i.e. pairs [First,Second] of tags of words occurring in the corpus adjacently and in this order;
- unigrams, i.e. the individual tags.

The qualitative representativity wrt. to bigrams consists in this case of two co mplementary parts:

- the representativity wrt. the presence of all possible bigrams of the language in the corpus, which means that if any bigram [First,Second] is a bigram in a correct sentence of the language, then such a bigram occurs also in the corpus — we shall call this positive representativity,
- the representativity wrt. the absence of all invalid bigrams of the language in the corpus, which means that if any bigram [First,Second] is a bigram which cannot occur in a correct (i.e. grammatical) sentence of the language, then such a bigram does not occur in the corpus — this we shall call negative representativity.

If a corpus is both positively and negatively representative, then indeed it can be said to be a qualitatively representative corpus[5] In our particular example this means that a bigram occurs in a qualitatively representative (wrt. bigrams) corpus if and only if it is a possible bigram in the language (and from this it already follows that any unigram occurs in such a corpus if and only if it is a possible unigram[6]).

[3] The case of a trigram-based tagger, more usual in practice, would be almost identical, but more lengthy. For the conciseness of argument, we limit the discussion to bigram-based taggers only, without loss of generality of argument towards general n-gram taggers.

[4] In an indeed broadly understood sense of the word "phenomenon" - of course a bigram is nothing that would be called a linguistic phenomenon (or at least not called a phenomenon in linguistics).

[5] The definitions of positive and negative representativity are obviously easily transferable to cases with other definitions of a phenomenon. Following this, the definition of qualitative representativity holds of course generally, not only in the particular case of corpus representative wrt. bigrams.

[6] This assertion holds only on condition that each sentence of the language is of length two (measured in words) or longer. Similarly, a corpus qualitatively representative wrt. trigrams is qualitatively representative wrt. bigrams and wrt. unigrams only on condition that each sentence is of length three at least, etc.

The requirement of quantitative representativity of a corpus wrt. bigrams can then be explained as the requirement that the frequency of any bigram and any unigram occurring in the corpus be "in exactly the right proportion" — i.e. in the proportion "as in the language performance" — to the frequency of occurrence of all other bigrams or unigrams, respectively. However, even when its basic idea is quite intuitive and natural, it is not entirely clear whether quantitative representativity can be formalized really rigorously. The main problem here is that what is at stake is measuring the occurrence of a bigram (or unigram, for that matter) within the full "language performance", understood as set of utterances of a language. This set, however, is infinite if considered theoretical ly (i.e. as set of all possible utterances in the language) and finite but practically unattainable if considered practically as a set of utterances in a language realized within a certain time span. Notwithstanding the theoretical problems, the frequencies are used in practice (e.g., for the purpose of training statistical taggers), and hence it is useful to have a look what they (the practical ones) really mean: in our example, it is the relative frequencies of the bigrams (and unigrams) in a particular (training or otherwise referential) corpus. However, nothing can be said as to whether this corpus is quantitatively representative or not in the original, intuitive sense.

3 Negative Representativity in Theory wrt. Bigrams

By quality we mean above all representativity wrt. the "phenomena" which the operation of the tagger is based on. (such as, e.g., bigrams or trigrams). In an ideal case this representativity would be a quantitative one. However, as shown above, the quantitative representativity is a problematic concept even in theory (let alone in practiced). For this reason, it seems more appropriate to strive (at least currently) for achieving a qualitatively representative corpus. In order to be able to achive this, we shall try to interpret the notion of bigram linguistically in this paragraph. The main aim of such an enterprise is to establish linguistic grounds for a corpus which is representative wrt. bigrams — and hence, to provide for means of creating such a corpus.

From a linguistic viewpoint, the pair of tags [First,Second] is a linguistically valid bigram in a certain natural language if and only if there exists a sentence (at least one) in this language which contains two adjacent words bearing the tags *First* and *Second*, respectively[7]. Such a sentence then can be assigned its (constituent) structure. This, in turn, means that for any valid bigram *[First,Second]* it is possible to find a level of granularity of the constituents such that this bigram is represented by a structural configuration where there occur two adjacent constituents *LC* (for "Left Constituent") and *RC* (for "Right Constituent"), such that LC immediately precedes RC and the last (rightmost) element of terminal yield of LC is First and the first (leftmost) element of the terminal yield

[7] As an example of a linguistically valid bigram of English, we can put forward, e.g., the pair *[finite_verb, adv]*. This bigram is valid because, e.g., the sentence *John walks slowly* is a correct sentence where the words *walks* and *slowly* bear the tags *finite_verb* and *adv*, respectively.

is Second, cf. Fig. 1, where also the common ancestor (not necessarily a direct ancestor!) of LC and RC is depicted (as AC, "Ancestor Constituent").

Fig. 1

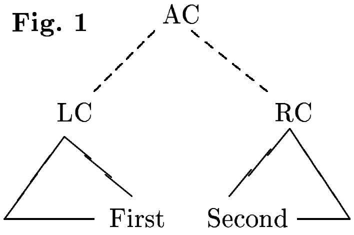
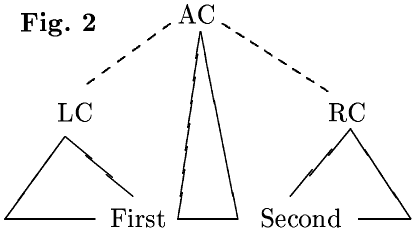


Fig. 2



Correspondingly, the pair of tags *[First,Second]* is a linguistically invalid bigram in a certain natural language if and only if there exists no grammatically correct sentence in this language which contains two adjacent words bearing the tags *First* and *Second*, respectively. Seen from a structural perspective, *[First,Second]* is an invalid bigram if one or more of the following obtains:

1. the configuration from Fig. 1 is impossible because in all constituents LC, *First* must necessarily be followed by some other lexical material[8];
2. the configuration from Fig. 1 is impossible because in all constituents RC, *Second* must necessarily be preceded by some other lexical material[9];
3. the configuration from Fig. 1 is impossible because LC and RC can never occur as adjacent sisters standing in this order — cf. Fig. 2[10].

4 Negative Representativity in Practice

In practice, the invalid bigrams occur in tagged corpora for the following reasons:

- in a hand-tagged corpus, an "invalid bigram" results from (and unmistakeably signals) either an ill-formed text in the corpus body or a human error in tagging;
- in a corpus tagged by a statistical tagger, an "invalid bigram" may also result from an ill-formed source text, and further either from incorrect tagging of the training data (i.e. the error was seen as a "correct configuration" in the training data, and was learned by the tagger) or from the process of so-called "smoothing", i.e. of assignment of non-zero probabilities also to

[8] Example: the bigram *[article,verb]* is impossible in English since in any LC — NPs, PPs, Ss etc. — article must be followed by (at least) a noun/adjective/numeral before an RC (in this case a VP or S) can start.

[9] Example: the bigram *[separable_verb_prefix,postposition]* is impossible in German since in any RC — NPs, PPs, Ss etc. — a *postposition* must combine with some preceding lexical material displaying (morphological) case before such a constituent can be combined with any other material into a higher unit.

[10] Example: the bigram *[finite_verb,finite_verb]* is impossible in Czech even when Czech has in the respect of verb position a completely free word order, and hence it would be possible that one finite verb stand at the end of LC and another one at the beginning of RC; however, in any AC — which in this case must be an S or a finite VP — the two finite verbs / verb phrases must be separated from each other by at least a conjunction (coordinating or subordinating) and/or by a comma.

configurations (bigrams, in the case discussed) which were not seen in the training phase[11].

However, the above theoretical considerations can be turned straightforwardly into a practical method for finding and removing the invalid (i.e. ungrammatical) bigrams in a standing corpus, achieving thus a high-quality PoS-tagging. The starting point for implementing this idea will be the search for the set of all invalid bigrams. For a particular language with a particular tagset, the set of invalid bigrams can be obtained by a reasonable combination of simple empirical methods leaning on the language performance that can be obtained from a corpus with a careful competence-based ("linguistic") analysis of the language facts. For the sake of easiness of explanation, let us make a provisional and in practice unrealistic assumption (which we shall correct immediately) that we have a qualitatively representative (wrt. bigrams) learning corpus of sentences of a certain language at our disposal[12]. Given such a (hypothetical) corpus, the next steps seem obvious and also easy to perform. First, all the bigrams in the corpus are to be collected to a set *VB* (valid bigrams), and then the complement of *VB* to the set of all bigrams (i.e. to the Carthesian product of the tagset) is to be computed; let this set be called *IB* (invalid bigrams). The idea is now that if any element of *IB* occurs in a PoS-tagged corpus whose correctness is to be checked, then the two adjacent corpus positions where this happened must contain an error (which then can be corrected). When implementing this approach to error detection, it is necessary to realize that learning the "invalid bigrams" is extremely sensible to the qualitative representativity of the learning corpus:

- the presence of an erroneous bigram in the set of *VB* causes that the respective error cannot be detected in the corpus whose correctness is to be checked (even a single occurrence of a bigram in the learning corpus means correctness of the bigram),
- the absence of a valid bigram from the *VB* set causes this bigram to occur in *IB*, and hence any of its occurrences in the checked corpus to be marked as a possible error (absence of a bigram in the learning corpus means incorrectness of the bigram).

However, the available corpora are neither error-free nor qualitatively representative. Therefore, in practice these deficiencies have to be compensated for by linguistic competence — by manual checking of the sets *VB* and *IB* obtained from the corpus.

5 Advanced Practice: "Stretching" the Invalid Bigrams

The "invalid bigrams" are a powerful tool for checking the correctness of a corpus, however, a tool which works on a very local scale only, and hence it is able to discover solely errors which are detectable as deviations from the set of possible

[11] This "smoothing" is necessary since — put very simply — otherwise configurations (bigrams) which were not seen during the learning phase cannot be processed if they occur in the text to be tagged.

[12] Note that we do not presuppose that the corpus is error-free — it might well contain tagging errors (and possibly other errors, e.g., ungrammaticalities in the input), only that none of them must cause an "invalid bigram" to occur in the corpus.

pairs of adjacently standing tags. Thus, obviously, quite a number of errors remain undetected by such a strategy. As example of such as yet "undetectable" errors we might put forward:

- the configuration article - adverb - verb in English,
- the configuration separable verb prefix - comma - postposition in German,
- the configuration finite verb - noun - finite verb in Czech.

There are two interesting observations to be done at these examples.

First, these configurations are wrong but they cannot be detected as such by the application of invalid bigrams: *[article, adverb]* and *[adverb, verb]* are certainly valid bigrams for English, and the same holds for the bigrams *[separable_verb_prefix, comma]* and *[comma, postposition]* in German, and also for the bigrams *[finite_verb, noun]* and *[noun, finite_verb]* in Czech.

Second, and more importantly, the linguistic argumentation why such configurations are erroneous is in fact the same as it was in the examples in Sect. 2:

- the English example configuration article - verb can become correct only if a non-pronominal nominal element (noun, adjective or numeral) is inserted inbetween the two elements, but not by insertion of an adverb (and, for that matter, neither by insertion of a preposition, conjunction, pronoun, verb,...)[13];
- likewise, the German example cannot be put right without insertion of a nominal element;
- and the Czech example will not be corrected until a comma or a conjunction is inserted inbetween the two finite verbs — any other material does not count for this purpose.

The central observation hence is that the property of being an impossible configuration is often retained also after the components of the invalid bigram get separated by certain kind of material occurring inbetween them. In fact, the addition of any amount of such material cannot make the configuration in question grammatical until the "right" material is inserted. Turned into practice, this observation yields a powerful tool for error detection in a corpus already tagged. In particular, it is possible to generalize on the set of already known invalid bigrams as follows. For each invalid bigram *[First,Second]*, collect all trigrams of the form *[First,Middle,Second]* occurring in the tagged corpus, and put all the possible tags *Middle* into the set *Allowed_Inner_Tags*. Further, given the invalid bigram *[First,Second]* and the set *Allowed_Inner_Tags*, search for all tetragrams *[First,Middle_1,Middle_2,Second]*. In case one of the tags *Middle_1*, *Middle_2* occurs already in the set *Allowed_Inner_Tags*, no action is to be taken, sice the tetragram is linguistically licensed by the legality of the respective trigram, but in case the set *Allowed_Inner_Tags* contains neither of *Middle_1*, *Middle_2*, both the tags *Middle_1* and *Middle_2* are to be added into the set *Allowed_Inner_Tags*. The same action is then to be repeated for pentagrams, hexagrams, etc., until the maximal length of sentence in the learning corpus prevents further prolongation

[13] This is not to say that by insertion of, e.g., a noun the configuration becomes necessarily correct — however, unless a nominal element is inserted, it remains necessarily incorrect.

of the n-grams and the process terminates. At last, construct the set *Impossible_Inner_Tags* as the complement of *Allowed_Inner_Tags* relatively to the whole tagset. Now, any n-gram consisting of the tag *First*, of any number of tags from the set *Impossible_Inner_Tags* and finally from the tag *Second* is very likely to be an invalid n-gram in the language. The respective algorithm in a semi-formal coating looks like as follows:

```
forall invalid_bigram [First, Second]
{
   n := 3;
   allowed_i_t := empty_set;
   while  n =< maximal_sentence_length_in_corpus
   do { find all inner-sentential n-grams
         [First, V1, V2, .., Vn-2, Second];
      for each n-gram found
      do if {V1, V2, .., Vn-2} n allowed_i_t = empty_set
         then allowed_i_t :=  allowed_i_t + {V1, V2, .., Vn-2};
      n := n + 1;
   };
   impossible_i_t([First, Second]) :=  tagset  -  allowed_i_t;
}
```

As above, this empirical (performance-based) result has to be checked manually (through a human language competence) for correctness, since the performance results might be distorted by tagging errors or by lack of representativity of the corpus.

6 Results on NEGRA

By means of the error-detection technique described above, we were able to correct 2.661 errors in the NEGRA corpus. The prevailing part of these errors was indeed that of incorrect tagging (only less than 8% were genuine ungrammaticalities in the source, about 26% were errors in segmentation). The whole resulted in changes on 3.774 lines of the corpus; the rectification of errors in segmentation resulted in reducing the number of corpus positions by over 700, from 355.096 to 354.354.

The paradoxical fact is, however, that the experience with research described above suggests strongly that large and simultaneously highly correct training data needed for statistical tagging cannot be achieved without using algorithms based on a good portion highly refined linguistic expertise.

7 Conclusions

The main contribution of this paper lies in the presentation of a method for detecting errors in part-of-speech tagged corpus which is both quite powerful (as to coverage of errors) and easy to apply, and hence it offers a relatively low-cost means for achieving high-quality PoS-tagged corpora. The main advantage is that the approach described is based on focussed search for errors of a specific type on particular "suspect" spots in the corpus, which makes it possible to detect errors even in a very large corpus where manual checking would not be feasible (at least in practice), since it requires passing through the whole of the text and paying attention to all kinds of possible violations.

The approach is also a very rewarding as to the possibilities of other applications it brings along with: in particular, it should not pass unnoticed that the set of invalid bigrams is a powerful tool not only for error detection in corpora already tagged, but also for avoiding errors in tagging raw texts, since an invalid bigram should never be used in — and hence never come into being as a result of — tagging a raw corpus(which, e.g., for a trigram-based tagger means that any trigram *[first,second,third]* containing an invalid bigram — i.e. if *[first,second]* or *[second,third]* are invalid bigrams — should be assigned probablity 0 (zero), and this also after smoothing or any similar actions are preformed). Another merit of the method worth mentioning is that it allows not for detecting errors only, but also for detecting inconsistencies in hand-tagging (i.e. differences in application of a given tagging scheme by different human annotators and/or in different time), and even inconsistencies in the tagging guidelines.

A particular issue is further the area of detecting and tagging idioms/collocations, in the case when these take a form deviating from the rules of standard syntax (i.e. they are detected as "suspect spots").

For details on all these points, including the particular problems encountered in NEGRA, cf. Květoň and Oliva (in prep.).

Acknowledgement

The quality of this paper profited immensely from comments supplied to us by Detmar Meurers. The work itself has been sponsored by the Fonds zur Förderung der wissenschaftlichen Forschung (FWF), Grant No. P12920. The Austrian Research Institute for Artificial Intelligence (FAI) is supported by the Austrian Federal Ministry of Education, Science and Culture.

References

1. NEGRA: www.coli.uni-sb.de/sfb378/negra-corpus
2. Brants T.: TnT - A Statistical part-of-speech tagger, in: Proceedings of the 6th Applied Natural Language Processing Conference, Seattle (2000)
3. Hirakawa H., K. Ono and Y. Yoshimura: Automatic refinement of a PoS tagger using a reliable parser and plain text corpora, in: Proceedings of the 18th Coling conference, Saarbrcken (2000)
4. Kveton P. and K. Oliva (in prep.) Correcting the NEGRA Corpus: Methods, Results, Implications, FAI Technical Report (in prep.)
5. Oliva K.: The possibilities of automatic detection/correction of errors in tagged corpora: a pilot study on a German corpus, in: 4th International conference "Text, Speech and Dialogue" TSD 2001, Lecture Notes in Artificial Intelligence 2166, Springer, Berlin (2001)
6. Schiller A., S. Teufel, C. Stckert and C. Thielen: Guidelines fr das Tagging deutscher Textcorpora, University of Stuttgart / University of Tbingen (1999)
7. Skut W., B. Krenn, T. Brants and H. Uszkoreit: An annotation scheme for free word order languages, in: Proceedings of the 3rd Applied Natural Language Processing Conference, Washington D.C. (1997)