POS Tagging without a Training Corpus or a Large-scale Lexicon: How far can one get?

António Horta Branco and João Ricardo Silva
{antonio.branco,jsilva}@di.fc.ul.pt
University of Lisbon, Faculty of Sciences

1 Introduction

Automatic part of speech (POS) tagging is a non-trivial procedure as it cannot be reduced to a simple lexical lookup process: For ambiguous types, those whose tokens admit POS tags that are different from each other, it is necessary to decide which tag to correctly assign in each occurrence of a corresponding token. From a broader, Natural Language Processing perspective, this tagging task is an efficiency-enhancing, local operation: The linguistic information uncovered is quite relevant for guiding and speeding up subsequent morphological and syntactic processing, but the clues taken into account in this decision procedure need only a narrow window of context, containing just a few tokens preceding or following the token to be tagged.

The mainstream methodologies for developing automatic taggers use data-intensive driven algorithms and techniques.[1] These techniques rely crucially on the existence of large training corpora previously tagged with full accuracy. Besides, these techniques may also benefit from the access to large-scale lexical resources where morphosyntactic information about lexical units is stored.

Both correctly tagged corpora and large-scale lexica are very expensive linguistic resources to develop and maintain, requiring a very large amount of work by trained teams of specialists. As a consequence, though the process of automatically developing an automatic tagger with good accuracy is now fast and easy, it presupposes previously prepared materials upon which it has to be trained and developed, materials that take a lot of hard effort, time and money to be constructed.

In the present paper we report on the results of an experiment we performed on POS tagging where no previously tagged corpus or large-scale lexicon were used. With this experiment we sought to obtain critical information on the following two aspects: (i) in terms of POS tagging, how far is it possible to get within the practical boundaries imposed by such constraints; and (ii) how useful can this strategy be in terms of supporting the task of obtaining a fully-fledged, correctly tagged corpus.

The data collected concerning (i) will allow us to establish a *measure of progress* inasmuch as we will be determining how much of a raw corpus can be in general automatically tagged without data-intensive tagging techniques or highly expensive linguistic resources. These results will be discussed in Section 2

[1] For instance, Brill (1992) used a rule transformation-based approach, Merialdo (1994) developed a tagger on the basis of hidden Markov models, and Ratnaparkhi (1996) proposed a maximum entropy-based tagger.

The information collected concerning (ii), in turn, will provide us with a *measure of benefit* inasmuch as we will be determining, with respect to the baseline task of completely and accurately hand tagging a corpus from scratch, how much of the effort required for this task can be avoided as it can be performed automatically with the low effort and inexpensive strategy put forward in (i). This aspect will be discussed in Section 3.

The present paper will be completed with Section 4, where we draw the conclusions from the experiment presented in the preceding sections.

2 Measure of progress

In order to proceed with our experiment, we developed a tagging tool that explores relevant linguistic regularities.

2.1 Rationale

From the typical distribution of frequencies of POS tags in a tagged corpus it is known that, for the few hundreds lexical types of the so-called morphosyntactic closed classes, their tokens exhibit very high frequencies.

We collected the list of items from closed classes, where we included every class other than Common Noun, Adjective, Verb, Adverb ending in *-mente*, Proper Name and Digit.[2] After having compiled this list of types from published grammars and online dictionaries, we developed a tagging tool that is able to tag the corresponding tokens by a process of simple lookup in that list.

As for the lexical items of the open classes referred to above, in a language like Portuguese, many of them are the result of productive morphological processes of word formation. As a rule, these processes and, more importantly, the category of the resulting words can be identified from their endings. We collected a list of this sort of endings and their categories, as well as the corresponding exceptions,[3] again from grammars and automatic searches in online dictionaries. With this in place, we made our tagging tool able to tag many of the tokens from open classes just by matching their endings with the ones in that list of ending-tag pairings.

To the above two procedures just described, we also added some simple heuristics that allowed our tool to also tag clitics and most proper names in the corpus.

Besides the tagging tool briefly described,[4] in order to perform our experiment, we also counted on two test corpora. One of these corpora was provided by CLUL-Centro de Linguística da Universidade de Lisboa,[5] which we will be calling in

---

[2] For the list of closed classes used, see the tag set in the Annex.

[3] For instance, *semente* is an exception to the rule that assigns ADV to tokens ending in *-mente*.

[4] For a detailed description of this tagging tool, see Branco and Silva (2002). In order to substantiate the reasoning throughout the present paper, it is important to recall that this tool was implemented so that if a type is handled by it, any occurrence of that type is ensured to receive all its grammatically admissible tags.

[5] We thank to Fernanda Bacelar and Amália Mendes for their kind help in this experiment.

the remainder of this paper as the CLUL corpus. This corpus has approximately 250 thousand tokens, resulting from the gathering of excerpts from newspapers, magazines, proceedings of meetings and novels.

The other test corpus we used is a portion of the CETEMPúblico corpus,[6] with approximately 11.5 million tokens, consisting of excerpts of the "Público" newspaper. The similarity of the results collected below for the two corpora suggests that these corpora are plausibly large enough so that distributional patterns of general relevance are uncovered within the boundaries of Zipfian expectations.

In the discussion that follows, there are three measurements we will take into account and whose definitions are:

- *Coverage of T:* N_T/C, where N_T is the number of tokens tagged with tag T, and C is the size of the corpus — this indicates how much of the whole corpus received the tag T, regardless of the fact that T coexists with other tags or not in the corresponding tokens.
- *Precision of T:* U_T/N_T, where U_T is the number of tokens tagged only with tag T, and N_T is the number of tokens tagged with T — for those tokens that received the tag T, this indicates how many of them received only this tag.
- *Progress for T:* U_T/C, where U_T is the number of tokens tagged only with tag T, and C is the size of the corpus — for the tag T, this indicates how much of the whole corpus received only this tag.

2.2 Closed classes

The table below displays the values obtained for both test corpora in terms of *coverage*, *precision* and *progress* for some prominent closed classes:

POS	Coverage		Precision		Progress	
	CLUL	CETEM	CLUL	CETEM	CLUL	CETEM
PREP	17.18%	18.46%	70.11%	71.31%	12.04%	13.16%
PNT	15.21%	14.10%	100.00%	100.00%	15.21%	14.10%
DA	12.74%	13.61%	37.93%	40.55%	4.83%	5.52%
CJ	6.55%	5.98%	33.52%	36.73%	2.20%	2.20%
IA	1.73%	1.71%	10.01%	10.27%	0.17%	0.18%
DEM	0.81%	0.71%	81.31%	74.36%	0.66%	0.53%
POSS	0.62%	0.61%	62.13%	51.43%	0.39%	0.31%
PRS	0.64%	0.22%	93.23%	87.06%	0.59%	0.20%

Table 1 - C*overage*, *precision* and *progress* for some closed classes

[6] http://cgi.portugues.mct.pt/cetempublico/

The largest *coverage* by a single category from the closed classes is ensured by Prepositions, with 17.18% – 18.46% of the whole corpus.

Together with Demonstratives, Possessives and Personal Pronouns, Prepositions exhibit a quite large value for *precision*, though the maximum value in this respect is obtained by Punctuation symbols, as expected.

In terms of *progress*, both Prepositions and Punctuation symbols show the highest scores, around 12% – 15%, each. In this respect, the remaining categories present much lower values: Either because they have a very small *coverage*, like Personal Pronouns, Possessives or Demonstratives; or because they have a very small *precision*, like Conjunctions or Indefinite Articles, which in the latter cases indicates that tokens tagged with these categories are very likely to end up also tagged with some other tag(s).

It is of note that the values are in general similar for both corpora. The larger fluctuations occur with the values for *coverage* and/or *precision* of Demonstratives, Possessives and Personal Pronouns. This can be explained because items from these categories are inflected for Person, and contrarily to the CETEM corpus, the CLUL corpus has text styles other than newspaper articles. Accordingly, items inflected for first and second person are proportionally more abundant in the CLUL corpus, and these are items that typically exhibit a lower degree of lexical ambiguity.

Taking into account the values in Table 1 and the values for the remaining closed classes, we obtained the following overall value of *progress* for closed classes:

	CLUL	CETEM
Progress	39.12%	39.38%

Table 2 - *Progress* for closed classes

2.3 Terminations

The table below shows the measures obtained for the categories that, for some of their tokens in the corpus, were identified by means of the terminations of those tokens.

POS	Coverage		Precision		Progress	
	CLUL	CETEM	CLUL	CETEM	CLUL	CETEM
ADJ	2.60%	2.76%	30.80%	31.55%	0.80%	0.87%
ADV (*-mente*)	0.44%	0.46%	98.36%	98.57%	0.44%	0.45%
CN	9.47%	9.92%	42.15%	47.37%	3.99%	4.70%
GER	0.33%	0.34%	97.95%	97.96%	0.32%	0.33%
PTP	2.37%	2.61%	58.23%	59.35%	1.38%	1.55%
V	10.11%	8.66%	74.35%	70.56%	7.52%	6.11%

Table 3 - C*overage*, *precision* and *progress* for classes detected by word terminations

As expected, the terminations with the highest *precision* are -*mente*, for Adverbs, and -*ando*, -*endo* and -*indo*, for the Gerund. Their overall *progress* is low, however, because there are few occurrences of these forms in both corpora, their *coverage* values presenting small values. This is in contrast with what can be observed for the terminations of Verbs: In spite of their lower value for *precision*, they are the most "useful", contributing more for the overall *progress*.

The *precision* values for Common Nouns and Adjectives are low because some terminations are associated with a *portmanteau* tag encompassing both of these categories, thus originating tokens ambiguously tagged.

The tokens correctly tagged via inspection of their terminations are around 14% of the test corpora.

	CLUL	CETEM
Progress	14.44%	14.34%

Table 4 - *Progress* for classes detected by word terminations

2.4 Heuristics

The following table presents the measures obtained for some categories that were identified by means of heuristics, to be described below:

POS	Coverage		Precision		Progress	
	CLUL	CETEM	CLUL	CETEM	CLUL	CETEM
PNM	7.04%	7.83%	97.75%	97.78%	6.88%	7.66%
CL	9.43%	9.14%	9.53%	6.44%	0.90%	0.59%
DGT	1.08%	1.44%	100.00%	100.00%	1.08%	1.44%
DGTR	0.13%	0.07%	91.50%	87.27%	0.12%	0.06%

Table 5 – C*overage*, *precision* and *progress* for some classes detected by heuristics

According to Portuguese orthography, Proper Names begin with a capital letter. We can take advantage of this fact to create a simple heuristic to identify some of their occurrences.[7] There are some exceptions that had to be taken into consideration, which usually also begin with a capital letter (e.g. social titles like *Presidente*, etc.), some of them being ambiguous between Proper Name and some other tag. Their existence is the reason why the Proper Name heuristic does not have a 100% value for *precision*.

Clitics in enclisis or mesoclisis are attached to the corresponding verb by a hyphen and are, therefore, easy to detect by a heuristic that takes into account this clue. However, when in proclisis, some Clitics have the same form as the Definite Articles that imposes that they receive more than one tag in such occurrences. Furthermore, most occurrences of Definite Articles, especially those not in a contracted form with

[7] Acronyms also begin with a capital letter and are handled by this heuristic.

some preceding preposition, are also annotated with the tag for Clitics. This set of circumstances explains why clitics present a low value for *precision*.

Digits and Roman Numerals are recognized by means of regular expressions. Roman Numerals do not have a 100% *precision* because some types can be ambiguous with some words, e.g. *vi*_V, *CML*_PNM, etc.

Adding the *progress* of each heuristic, we obtain the following values for the overall *progress* of the heuristics we used.

	CLUL	CETEM
Progress	8.98%	9.83%

Table 6 - *Progress* for classes detected by heuristics

2.5 Global results

Given the similarity of global results for the two test corpora in the preceding sections, for the sake of perspicuity, we took only one of these corpora, namely the CETEM corpus, to build the following chart:

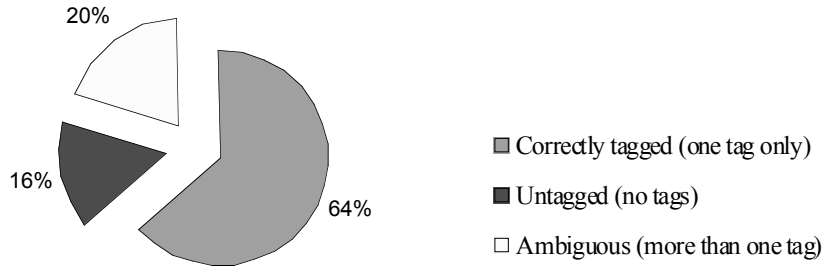

Chart 1 – Results for the CETEM corpus:
16% received no tag, 64% one tag only, and 20% more than one tag

This shows that, when doing automatic POS tagging without a training corpus or a large-scale lexicon, the value of *progress* can reach as much as 64% — this is the portion of the corpus that is already correctly tagged — leaving only 36% to be concluded.

These 36% of the corpus can be further divided into tokens that received no tag yet, which amount to 16% of the corpus, and tokens that were annotated with more than one tag, 20% of the corpus.

3 Measure of benefit

Having obtained a measure of progress, we focus now on estimating a measure of benefit for the partial tagging method we used.

Given the particulars of this method, if a token ended up annotated with more than one tag, then it bears all its admissible tags. Accordingly, the tag to be finally assigned to each of these tokens is to be selected from the set of tags already assigned to it.[8] To the portion of the corpus made of tokens in this circumstance, that received more than one tag, we called the *detected ambiguity*. It amounts to approximately 20% of a corpus, as displayed in the chart above.

On the other hand, to the portion of tokens in the corpus that received no tag, we called *remaining ambiguity*: Each of these still untagged tokens can be considered as being ambiguous over four tags: Common Noun, Adjective, Verb and Part of Name, inasmuch as its final tag is certainly one of these four. Remaining ambiguity is 16% of a corpus.

3.1 Resolving ambiguity

With respect to the *detected ambiguity*, in view of evaluating the magnitude of the corresponding disambiguation task, it is interesting to consider the degree of ambiguity involved. We recorded the distribution of tokens with different degrees of ambiguity, i.e. the distribution of tokens with different number of tags concomitantly assigned to them. The results were compiled in the following table:

Ambiguity	Distribution	
	CLUL	CETEM
2 tags	58.96%	59.23%
3 tags	29.71%	29.14%
4 tags	0.56%	0.63%
5 tags	0.09%	0.09%
6 tags	1.64%	1.58%
MWU vs. parts	9.04%	9.34%

Table 7 – Distribution of the *detected ambiguity*

The table shows that most of the *detected ambiguity* involves only 2 tags per token. As the number of tags per token increases, the frequency of such ambiguities decreases. There is, however, an odd increase in the frequency of ambiguities involving 6 tags. This is due to the words *como* and *nada*. Both are very frequent and ambiguous.[9]

---

[8] See footnote 4.

[9] *como* receives the tags INT, REL, CJ, PREP, ADV, and V, and occurs 33,291 times while *nada* receives the tags IN, DIAG, ADV, CN, ADJ, and V, and occurs 3,771 times.

MWU vs. parts is a special case of ambiguity where the tokens in a sequence of tokens can be tagged collectively as a multi-word unit or individually.

The values for frequencies in the previous table, can be used to get the values of *coverage* for the different degrees of ambiguity in the corpus, summarized as follows:

Ambiguity	Coverage	
	CLUL	CETEM
2 tags	12.47%	12.07%
3 tags	6.28%	5.94%
4 tags	0.12%	0.13%
5 tags	0.02%	0.02%
6 tags	0.35%	0.32%
MWU vs. parts	1.91%	1.90%

Table 8 – *Coverage* of the degrees of ambiguity

3.2 Comparing with a baseline

In order to have some sensible *measure of benefit* of the tagging approach described in this paper, one needs to contrast the amount of effort required by this approach with a baseline, which in the present case is the amount of effort required by hand tagging a raw corpus from scratch.

If one is left with less tokens to be tagged, as it happens in the former case, then that is already an indication that this approach represents a benefit for the task of accurately tagging a corpus. But we would like to have a more sophisticated method of measuring the benefit of our approach. In particular, we would like to accommodate in the final measure of benefit also the improvement resulting from having circumscribed the detected ambiguity and the remaining ambiguity. In this respect, the intuition that seems reasonable to account for is that:

(i) deciding which tag to assign to a token in the *detected ambiguity* part of the corpus is easier than deciding which tag to assign to that token if this were to be done from scratch — in the latter case, one has to decide which tag to choose and assign from the whole tag set (i.e. typically from around 40 – 70 tags), while in the former case one has to decide which tag to choose from a much more restricted set of tags, namely those that were already automatically assigned to that token, with not more than a few tags (2 – 6 in our experiment);

(ii) deciding which tag to assign to a token in the *remaining ambiguity* part of the corpus is easier than deciding which tag to assign to that token if this were to be done from scratch — in the former case one has to decide which tag to choose from a set with only four tags, namely the tags of the classes Common Noun, Adjective, Verb and Part of Name.

We are aware that building a fully accurate model of the effort involved in the human decision/tagging process would involve a whole set of performance variables whose detection and evaluation is clearly outside the scope of an experiment like the present one. Nevertheless, we think it is possible to build a first reasonable approximation to it on the basis of some basic counting.

The effort model we propose below represents a first contribution in this direction, to be certainly improved by subsequent research, but that already provides a first glimpse into the magnitude of the effort involved or saved.

3.3 A first measure of benefit

In a facilitating tool for hand tagging like EtiFac (Branco and Silva, 2002), after a token to be tagged has been automatically detected and selected, the user only has to scan through a list of tags in a drop down menu and choose which tag to be assigned to that token. Given this kind of decision procedure for human taggers, the amount of effort involved in a tagging decision is directly proportional to the number of tags that have to be inspected for the decision to be made.

Taking a list of N possible tags, we assume that on average, one needs (N+1)/2 "inspection" steps to find the desired tag: We will thus consider that, in the context of tagging a large enough corpus, the effort for tagging a token is proportional to (N+1)/2 with a tag set of size N.

For the sake of concreteness, let us take the CETEM corpus and a tag set with 39 tags. We can now estimate a magnitude for the effort required to hand tag the CETEM corpus from scratch:

- There are 11,523,947 tokens to be tagged;
- On average, the steps required to tag each token are (40 / 2 =) 20;

Therefore, globally, the number of steps required to completely tag the corpus is around (11,523,947 x 20 =) 230,478,940. This provides affair approximation to a baseline value.

It is also possible to estimate a magnitude for the effort required to tag the CETEM corpus after it having been partially tagged by the tagging device described above. In this case, we have to consider two possible circumstances for the tokens that have yet to be tagged:

- *Remaining ambiguity* — There are 1,894,594 tokens in the remaining ambiguity part of the corpus, and four possible tags for each token (CN, ADJ, V and PNM). The number of steps required for tagging this part of the corpus can be approximated as (1,894,594 x **2.5** =) 4,736,485.
- *Detected ambiguity* — There are 2,347,944 tokens in the detected ambiguity part of the corpus. The number of tagging steps should be calculated now in accordance with the frequency of the different degrees of ambiguity (vd. Table 9): This turns out to yield (1,390,578 x **1.5** + 684,089 x **2** + 14,837 x **2.5** + 2,115 x **3** + 37,062 x **3.5** + 219,263 x **2** =) 4,065,726.

Tags	Freq.	Weight	Effort
2 tags	1,390,578	1.5	2,085,867
3 tags	684,089	2.0	1,368,178
4 tags	14,837	2.5	37,093
5 tags	2,115	3.0	6,345
6 tags	37,062	3.5	129,717
MWU vs. Parts	219,263	2.0	438,526

Table 9 – Frequency for the degrees of detected ambiguity

Adding the values for these two possible circumstances, the total tagging steps can be approximated as 8,802,211.

With the above two values, 230,478,940 with respect to the effort of tagging the corpus from scratch, and 8,802,211 with respect to the effort of tagging the corpus with the help of our tagging device, the *measure of benefit* can be calculated as 1 - (8,802,211 / 230,478,940 =) 0.9618. The effort effectively saved by the tagging approach described in this paper is thus estimated as being around **96.18%** of the effort that would have been required in case one had tagged the corpus from scratch: For hand tagging a corpus, only ca. 4% of the original effort/resources is now required.

4 Conclusions

In this paper we described an experiment that permitted us to determine how far is it possible to get, in terms of POS tagging, without using a training corpus or a standard, large-scale lexicon. By exploring known ratio types/tokens for closed classes items, fairly invariant across languages and genres, word terminations regularly associated with specific POS tags and corresponding exceptions, and straightforward heuristics for numerals and proper nouns, it is possible to rapid prototyping a facilitation tool for POS tagging. When run over different corpora, this tool was shown to be able to accurately tag ca. 64% of a corpus.

With this experiment we sought also to determine how useful can this strategy be in terms of reducing the effort of obtaining a fully-fledged, accurately tagged corpus. When taking into account not only the ca. 64% accurately tagged, but also both the ambiguous tagging exhaustively provided and the reduction of the decision space for the untagged tokens in the remainder 36%, one saves ca. 96% of the effort/resources that would be required by a baseline tagging procedure consisting of hand tagging each and every token in a raw corpus.

It is certainly possible to devise a more accurate model of the human tagging process, more sophisticated than the one used here. Building such model would involve a whole set of performance variables whose detection and evaluation is clearly outside the scope of a paper such as the present one. Nevertheless, we think that the results reported here represent a first contribution in this direction, to be improved by

subsequent research,[10] but that already provide a first glimpse on the magnitude of the effort involved or saved.

5 References

Branco, António and João Silva, 2002, "EtiFac: A Facilitating Tool for Manual Tagging". In *Actas do XVII Encontro Anual da APL*, pp. 81 – 89.

Branco, António and João Silva, 2003, "Accurate Annotation of Corpora: How Efficient is the Most Efficient Method?", ms., University of Lisbon, Faculty of Sciences.

Brill, Eric, 1992, "A Simple Rule-Based Part-of-Speech Tagger", In *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing*, Trento, IT, pp.152-155.

Merialdo, B., 1994, "Tagging English Text with a Probabilistic Model". *Computational Linguistics*, 20, pp. 156 – 171.

Ratnaparkhi, Adwait, 1996, "A Maximum Entropy Model for Part-of-Speech Tagging", In Eric Brill and Kenneth Church (eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ACL, pp. 133 – 142.

[10] For steps towards refining the measure of benefit presented here, see Branco and Silva (2003).

Annex

Tag	Category	Examples
ADJ	Adjective	*bom, brilhante, eficaz*
ADV	Adverb	*hoje, já, sim, felizmente*
CARD	Cardinal	*zero, dez, cem, mil*
CJ	Conjunction	*e, ou, tal como*
CL	Clitic pron.	*o, lhe, se*
CN	Common noun	*computador, cidade, ideia*
DA	Def. article	*o, os*
DEM	Demonstrative	*este, esse, aquele*
DFR	Denominator of fraction	*meio, terço, décimo, %*
DGT	Digit	*0, 1, 42, 12345, 67890*
DGTR	Roman numeral	*VI, LX, MMIII, MCMXCIX*
DIAG	Dialogue particle	*adeus, olá, alô*
EADR	Electronic address	*http://www.di.fc.ul.pt*
EOE	End of enum.	*etc*
GER	Gerund	*sendo, afirmando, vivendo*
IA	Indef. article	*um, uns*
IN	Indefinite nominal	*tudo, alguém, ninguém*
INT	Interrogative pron.	*quem, como, quando*
ITJ	Interjection	*oh, ah, eh*
LTR	Letter	*a, b, c*
MGT	Magnitude	*unidade, dezena, dúzia, resma*
MTH	Month	*Janeiro, Dezembro*
NP	Noun phrase	*idem*
ORD	Ordinal	*primeiro, centésimo, penúltimo*
PADR	Part of address	*rua, av., rot.*
PNM	Part of name	*Lisboa, António, João*
PNT	Punctuation marks	*., ?, (*
POSS	Possessive	*meu, teu, seu*
PP	Prepositional phrase	*algures*
PREP	Preposition	*de, para, em redor de*
PRS	Personal pron.	*eu, tu, ele*
PTP	Past participle	*sido, afirmado, vivido*
QD	Quantifier det.	*todos, muitos, nenhum*
REL	Relative pron.	*que, cujo, tal que*
STT	Social title	*Presidente, drª., prof.*
SYB	Symbol	*@, #, &*
TERMN	Optional terminations	*(s), (as)*
UNIT	Unit of measure	*km, kg, b.p.m.*
V	Verb (other than PTP or GER)	*ser, afirmar, viver*
WD	Week day	*segunda, terça-feira, sábado*