Are Morphosyntactic Taggers Suitable to Improve Automatic Transcription?

No Author Given

No Institute Given

Abstract. The aim of our paper is to study the interest of part of speech (POS) tagging to improve speech recognition. We first evaluate the part of misrecognized words that can be corrected using POS information; the analysis of a short extract of French radio broadcast news shows that an absolute decrease of the word error rate by 1.1% can be expected. We also demonstrate quantitatively that traditional POS taggers are reliable when applied to spoken corpus, including automatic transcriptions. This new result enables us to effectively use POS tag knowledge to improve, in a postprocessing stage, the quality of transcriptions, especially correcting agreement errors.

1 Introduction

Automatic speech recognition (ASR) systems globally use little knowledge about language. Very often, linguistic knowledge is limited to the learning of probabilities of word sequences on a corpus. However, as ASR systems act on natural language, linguistic knowledge should improve the quality of transcription. Some language models (LM) already integrate linguistic knowledge such as syntactic structures of utterances [1], topics of the document to transcribe [2, 3] or parts of speech (POS) [4]. A part of speech is a grammatical property of a word, or a group of words, in a given sentence (*e.g.* nouns, verbs, prepositions, conjunctions, *etc.*) often along with morphological information (gender, number, conjugation, *etc.*). The knowledge of these categories is generally included in LM thanks to class-based N-gram models [5]. Formally, if we denote by $\mathcal{C}_i$ the set of possible POS tags for a word w_i, the probability of the word sequence $w_1^n = w_1, ..., w_n$ is obtained as follows:

$$P(w_1^n) \approx \sum_{c_1 \in \mathcal{C}_1, ..., c_n \in \mathcal{C}_n} \prod_{i=1}^{n} P(w_i|c_i) P(c_i|c_{i-N+1}^{i-1}) \quad (1)$$

The interpolation of a class-based N-gram models with a word-based N-gram model generally results in a negligible decrease of the word error rate (WER) of transcription and different improvements have been proposed. For example, probability can be estimated by considering that the POS tags of the words w_i to recognize are directly incorporated in the production of the ASR systems and are not only an intermediate result [6]. This approach evaluates the probabilities

thanks to more accurate calculations than in (1) but leads to a dramatic increase of the number of events to consider.

In all these approaches, POS, *i.e.*, morphosyntactic information allow to build LMs which are further used in the transcription process. As experiments have resulted in a limited improvement with respect to word-based N-gram LMs, we propose to investigate on the use of POS tags in a postprocessing stage to correct some errors. In the first part of our work, we study automatic transcriptions and show that the proportion of errors that could be corrected by the knowledge of POS information is significant. We then demonstrate the ability of automatic taggers to deal with spoken language and with transcription errors. We finally evaluate several approaches to use POS tag information to rescore the N-best sentence candidates produced by our ASR system. Our experiments show that a majority of gender and agreement errors are corrected. However, new errors are also introduced, thus resulting in a marginal decrease of the WER.

2 Typology of transcription errors

In order to evaluate the potential contribution of POS taggers for automatic speech transcription, we first analyze closely a short excerpt of an automatic transcription in order to measure the proportion of errors that can be corrected by POS knowledge.

The ASR system used in our experiments, designed for the transcription of broadcast news shows in the French language, aims at producing a word graph using a three pass strategy. In a first pass, a large word graph is generated using context-independent acoustic models and a trigram LM. The second pass aims at rescoring the 1,000 best paths of the first pass word graph after expansion by a 4-gram LM. Finally, the third pass is similar to the second one but uses speaker adapted acoustic models, where MLLR adaptation is carried out using the second pass output. A smaller word graph is also generated by the third pass, with about 80 word boundaries per second. To get rid off segmentation problems, we consider in this work a manual segmentation by utterances, where an utterance is actually a breath group. On the ESTER [7] broadcast news corpus, the overall word error rate on the entire corpus is 22.4 % while the word graph error rate is 13.9 %. However, in this section, we analyze the transcription errors on a subset of the entire corpus, where the WER is 17.8 %.

Among the transcription errors we observed, three groups stand out. Some errors are caused by a "drift" of the ASR system, generally explained by either a bad acoustics or a misrecognition of named entities which are out of vocabulary. These errors seem to be out of the scope of POS-based techniques. Fortunately, they only affect a restricted part of the analyzed extract. The second set of errors is related to ungrammatical transcriptions (Fig. 1) which can in particular be caused by short grammatical words like "*a*" ("*has*"), "*à*" ("*at*"), "*de*" ("*of*"), or "*et* ("*and*"), that are sometimes missing or wrongly inserted in the transcription hypotheses. We also find tense and mood errors for verbs, with a too systematic prevalence for the present indicative tense. Among these errors, some are

```
REF: bush ** SAIT donc QU' il faudra coopérer
HYP: bush s' DONC DONC **  il faudra coopérer
```

Fig. 1. Ungrammatical utterance

```
REF: c' est un monstre injuste envers sa soeur si DÉVOUÉE
HYP: c' est un monstre injuste envers sa soeur si DÉVOUÉ
```

Fig. 2. Example of agreement error

rectifiable using POS knowledge since the tagging of utterances can produce absurd POS sequences, such as three consecutive prepositions. Nevertheless, this criterion must be cautiously used, because of repetitions which naturally appear in spoken language. The third set represents errors most probably rectifiable by POS, *i.e.*, erroneous gender and number agreements. These errors are particularly numerous, affecting a seventh of the utterances. Most of these errors come from the fact that, in the French language, the plural and singular forms, or the masculine and feminine forms of many words are homophone. There are also many homophone confusion between the various tenses of verbs. Among these errors, 70 are rectifiable without inspecting dependencies between consecutive utterances (Fig 2); their correction would result in an absolute decrease of 1.1 % of the WER. Through the report of the main decoding errors, it therefore appears that POS knowledge is a valuable information to improve transcription quality.

3 Behavior of taggers

The previous section showed the interest of POS taggers to correct some transcription errors by focusing on the possible sequences of POS. However, to use this technique, taggers must reliably operate on spoken corpora, produced by annotators or by ASR systems. It is this property that is evaluated in this section.

Morphosyntactic taggers aim at associating the most likely tag with each word or word group, in the word sequence to study. These tools are normally applied on written corpora. To quantitatively evaluate them, a text is manually tagged by annotators and tags are compared one by one with those automatically produced. In comparison to written corpora, spoken corpora transcribed by human listeners have been seldom studied [8]. Oral production has however characteristics, such as repetitions, revisions or fillers, that may complicate tagging. POS tagging of the automatic transcription of planned speech is an even more complex task as the text is segmented in breath groups rather than in sentences, and lacks punctuation and, in the case of our ASR system, capital letters.

In order to make the use of POS easier to decode speech, we decided to build our own morphosyntactic tagger. The remainder of this section first describes the protocol used to build this tagger, before evaluating its behavior on transcribed speech. We measure the quality of the tagging output on a test corpus and compare the results with those obtained by a standard tagger for the French language.

3.1 Tagger design

The taggers conceived for written documents use linguistic rules or automatically extract statistic information from voluminous data. Given that programs based on statistical calculations produce satisfactory results for written documents and do not require to manually write numerous contextual rules, we built our tagger by solely using statistical methods. Another reason for this choice is the ability of statistical taggers to provide scores for a sequence of tags.

With this intention in mind, we established a 200,000-word training corpus representing a 16-hour extract from the ESTER corpus. The manual transcriptions, originally containing capital letters and punctuations, were tagged by the Cordial software [1]. We manually checked the result, before removing all capital letters and punctuation marks to obtain a format similar to the one of the text produced by our ASR system. We used a tagged pronunciation lexicon to know all the possible POS for each word. We chose our morphosyntactic tags to distinguish the gender and the number of adjectives and nouns, and the tense and the mood of verbs, which led to a set of 80 tags. This set is very similar to the ones proposed in school grammars and directly inspired by Cordial's one.

Our morphosyntactic tagger is based on a class-based N-gram model to find the tags sequence:

$$\hat{c}_1^n = \arg\max_{c_1^n} \prod_{i=1}^{n} P(w_i|c_i) P(c_i|c_{i-N+1}^{i-1}) \tag{2}$$

for a word sequence w_1^n. Adjustments on a development corpus led us to choose a $N = 7$ order and an unmodified Kneser-Ney smoothing. To evaluate the effect of segmentation on the quality of tagging, we proceeded with two trainings, by segmenting the training corpus by sentence and by utterance.

3.2 Tagging evaluation

To get a quantitative measure of the quality of tagging for manually produced transcriptions (REF), segmented by sentence or utterance, or for transcriptions automatically produced by the decoder (HYP), we manually tagged a 1-hour broadcast news from the French radio station France-Inter. This broadcast, incorporating 11,300 words, will be referred to as GOLD in the sequel. The automatic tagging of REF was evaluated by counting the number of shared tags

[1] Distributed by the *Synapse Développement* corporation.

Table 1. Evaluation of taggers (in percentages)

	REF / sentence	REF / utterance	HYP
training corpus / sentence	91.42	91.09	72.60 (91.83)
training corpus / utterance	91.50	91.42	72.99 (92.32)
Cordial	88.69	88.61	70.75 (89.48)

with GOLD. The measure of the quality of tagging was problematic for HYP, for which we measured a WER of 22 %, since words differ from those of GOLD. It was thus impossible to form a reference tagging for HYP as there does not exist any valid POS for the words of ungrammatical utterances. We give two measures of the quality of the tagging of HYP: the percentage of correctly recognized and tagged words among all the words of GOLD, and the percentage of correctly recognized and tagged words among all the correctly recognized words in HYP (this latter number is given in brackets in Table 1).

The results obtained by our tagger on the test corpora are given in the two first lines of Table 1, by combining all the possible segmentations of the training and test corpora. These results establish that tags are globally correct, including for the automatic transcriptions for which recognition errors were likely to jeopardize the tagging of correctly recognized words. This is quite surprising since we did not resort to specific methods to deal with the particularities of spoken language, apart from the use of an oral corpus to estimate the tagger probabilities. However, this robustness is due to the fact that taggers locally assign possible tags and that many words are unambiguous. Therefore, even if the transcription is partially erroneous, unambiguous words correctly recognized helps keeping the tagger on tracks. Besides, the results show that training from the segmentation by utterance produce the best results, which led us to prefer this kind of segmentation later on.

Besides, by inspecting errors done during the assignation of POS, we noticed some could be considered as acceptable. For example, the distinction between the POS "past participle" and "adjective" are in a high majority of cases questionable. We also observed that numerous errors are caused by a wrong tokenisation of our tagger. For instance, while "*états-unis*" ("*united states*") was tagged as a proper name in GOLD, automatic tagging led to recognize on the one hand "*états*" ("*states*") as a noun and on the other hand "*unis*" ("*united*") as an adjective. Among the 966 errors observed for the tagging of REF segmented by utterance, 42 were caused by confusions between past participle and adjective, 216 by wrong tokenisation, 124 by confusions between common nouns and proper names and 10 by words unknown by the tagger.

Finally, we compared the performances of our tagger with those of Cordial, probably the best tagger available for written French, which has already produced good results on a spoken corpus [8]. The last line of table 1 shows the results. Our tagger yields results comparable to those obtained with Cordial, and even better. On this particular corpus, the Cordial tagger does perform poorly compared to results obtained with it on written document, usually above 95 %

```
REF: la vie politique TOUTE entière A ÉTAIT DÉSTABILISÉE
COR: la vie politique TOUTE entière * EST   DÉSTABILISÉE
HYP: la vie politique TOUT  entière * EST   DÉSTABILISÉ
```

Fig. 3. Example of an utterance where transcription agreement errors were manually corrected

correct. This is explained by the particularities of the automatic transcriptions, for which Cordial was not specifically conceived. The lack of capital letters is particularly problematic since Cordial relies on this information to detect proper names. By ignoring all the errors caused by confusion between proper names and common names, the percentage of correctly assigned tags rises to 93.52 % for the test corpus segmented by utterance, while, according to the same criterion and on the same test data, the performances of our tagger only improve up to 92.55 %.

This sequence of experiments shows that the tagging of automatic transcription *is* reliable, assertion which was only a hypothesis before. Our tagger leads to results that allow us to use it to score the quality of the decoding. In the following section, we present experiments on using POS knowledge in a postprocessing step to reevaluate N-best candidate sentences.

4 Contribution of tagging to transcription

To exploit the knowledge of POS during speech decoding, we used our tagger to score each hypothesis found for an utterance. For each candidate sentence w_1^n, the most likely corresponding tag sequence c_1^n is computed according to (2) before evaluating a score function given by

$$\mathrm{lp}(w_1^n) = \log P(c_1^n) = \sum_{i=1}^{n} \log P(c_i | c_{i-N+1}^{i-1}) \ . \tag{3}$$

We also used a normalized version of the score function, $\mathrm{lpn}(w_1^n) = \mathrm{lp}(w_1^n)/n$. Theses scores aim at reordering the list of the hypotheses produced by the ASR system for each utterance.

To validate our approach, we first tested the behavior on the 70 agreement errors found on the limited corpus analyzed in section 2. For each utterance containing one of these errors, we evaluate the score function (3) on three versions of the utterance: the reference transcription (REF), the automatic transcription (HYP) and the automatic transcription where agreement errors have been manually corrected (COR). Figure 3 shows the three version for an example utterance. The motivation for this first experiment is to verify that the POS score function is able to correctly rank the three versions of the utterance, with the higher score on REF and the lower on HYP. We observed that, for the 63 analyzed utterances, scores were higher on COR than on HYP for 46 of them

with lp and 47 with lpn, and were higher on REF than on HYP for 41 with lp and lpn. These results establish therefore that, in a majority of cases, both scores allow to correct agreement errors and are likely to reduce the WER.

We then used the POS score functions to reorder the list of 100-best hypotheses produced for 4 hours of French broadcast news. The 100-best list has an oracle word error rate of 14.2 % while the initial WER is 22.0 %. When reranking the N-best lists based only on POS scores, the WER increases significantly from 22.0 % to 26.2 % with lp and 27.5 % with lpn. Therefore, we decided to combine the non normalized POS score lp, which seemed to have a better behavior in the previous experiment, with the acoustic score and the LM score.

ASR is, in practice, usually expressed as a search of w_1^n from the acoustic input y_1^n, *i.e.*,

$$\hat{w}_1^n = \arg\max_{w_i^n} \; logP(y_1^n|w_1^n) + \alpha \; logP(w_1^n) + \gamma \; n \tag{4}$$

where α is the LM scale factor and γ a word insertion penalty term. To introduce POS scores, we extended the maximization (4) to

$$\hat{w}_1^n = \arg\max_{w_i^n} \; logP(y_1^n|w_1^n) + \alpha \; logP(w_1^n) + \beta \; logP(c_1^n) + \gamma \; n \; . \tag{5}$$

$P(w_1^n)$ is computed by a 4-gram LM on words while $P(c_1^n)$ is determined by a 7-gram LM on POS tags.

We observed a slight decrease of the WER down to 21.4 % with this method. Figure 4 reports the WER without the POS score (*i.e.* $\beta = 0$) and with the POS score after optimization of β. In both cases, γ is fixed for α and β to enable the lowest possible WER. The results show that for all α, tagging score leads to a very slight but consistent decrease of the WER. We generally noticed that agreement errors were corrected. For instance, an utterance transcribed as "*le messin disputent aujourd'hui*" ("*the inhabitant of Metz play today*") was correctly rectified in "*le messin dispute aujourd'hui*" ("*the inhabitant of Metz plays today"*) but a few new errors appear like the correct transcription "*les visages de Jacques Chirac et Jean-Marie Le Pen apparaissent*" ("*the faces of Jacques Chirac and Jean-Marie Le Pen appear*") erroneously rectified in "*les visages de Jacques Chirac et Jean-Marie Le Pen apparaît*" ("*the faces of Jacques Chirac and Jean-Marie Le Pen appears*").

POS knowledge brings restricted information in relation to word-based LM, although both methods are complementary as the consistent reduction of the WER proves it.

5 Future works

In this paper, we have shown that a significant proportion of transcription errors for the French language concerns agreement errors and can be corrected by POS knowledge. We have quantitatively proved that taggers can be used on spoken corpora transcribed by annotators or obtained by ASR systems. This allows us

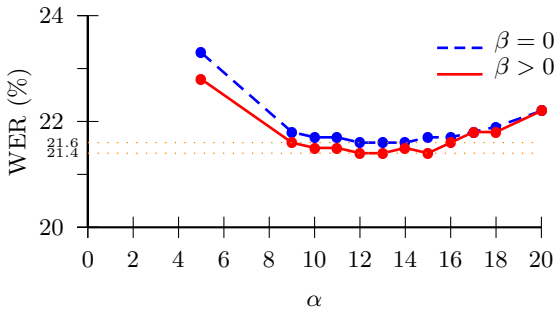

Fig. 4. WER as a function of the LM scale factor α, with ($\beta > 0$) and without ($\beta > 0$) POS information

to exploit POS tagging to improve transcriptions. Our experiments show that POS taggers are suitable to correct some agreement errors, even if it globally only results in a slight decrease of the WER. To better those first results, instead of operating on the N-best hypotheses produced by an ASR system, we plan to rescore all the homophones of the best hypothesis found [9]. Besides, we want to investigate other sets of POS tags for our tagger.

References

1. Chelba, C., Jelinek, F.: Structured language modeling. Computer Speech and Language **14**(4) (2000) 283–332
2. Khudanpur, S., Wu, J.: A maximum entropy language model to integrate n-grams and topic dependencies for conversational speech recognition. In: Proc. of ICASSP. (1999)
3. Iyer, R., Ostendorf, M.: Modeling long distance dependence in language: Topic mixtures versus dynamic cache models. IEEE Transactions on Speech and Audio Processing **7**(1) (1999) 30–39
4. Maltese, G., Mancini, F.: An automatic technique to include grammatical and morphological information in a trigram-based statistical language model. In: Proc. of ICASSP. (1992)
5. Brown, P., Della Pietra, V., deSouza, P., Lai, J., Mercer, R.: Class-based n-gram models of natural language. Computational Linguistics **18**(4) (1992) 467–480
6. Heeman, P.: POS tags and decision trees for language modeling. In: Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. (1999)
7. Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.F., Gravier, G.: The ESTER phase ii evaluation campaign for the rich transcription of French broadcast news. In: Proc. of Eurospeech. (2005)
8. Valli, A., Véronis, J.: Étiquetage grammatical de corpus oraux : problèmes et perspectives. Revue française de linguistique appliquée **4**(2) (1999) 113–133
9. Gauvain, J.L., Adda, G., Adda-Decker, M., Allauzen, A., Gendner, V., Lamel, L., Schwenk, H.: Where are we in transcribing French broadcast news? In: Proc. of Eurospeech. (2005)