Morphosyntactic Processing of N-Best Lists for Improved Recognition and Confidence Measure Computation

Stéphane Huet, Guillaume Gravier, Pascale Sébillot

IRISA
Campus de Beaulieu, F-35042 Rennes Cedex, France
{shuet,ggravier,sebillot}@irisa.fr

Abstract

We study the use of morphosyntactic knowledge to process N-best lists. We propose a new score function that combines the parts of speech (POS), language model, and acoustic scores at the sentence level. Experimental results, obtained for French broadcast news transcription, show a significant improvement of the word error rate with various decoding criteria commonly used in speech recognition. Interestingly, we observed more grammatical transcriptions, which translates into a better sentence error rate. Finally, we show that POS knowledge also improves posterior based confidence measures.

Index Terms: speech recognition, parts of speech, confidence measures

1. Introduction

Automatic speech recognition (ASR) systems globally use little knowledge about language. This is explained by the limited success met by techniques introducing new linguistic features, by comparison with word-based language models (LMs). In particular, parts of speech (POS), which are grammatical classes (*e.g.* verb, noun, preposition...) of words or locutions, are often considered as poorly informative to improve LMs. However, a significant number of transcription errors could be corrected by information about gender and number agreement [1]. In particular, in the French language, nouns, adjectives, and verbs are very often inflected for number, gender or tense into various homophone forms; this makes POS, *i.e.*, morphosyntactic information, interesting to improve the quality of transcription, at least for highly inflectional languages. Moreover, morphosyntactic knowledge can be reliably and automatically deduced by taggers, even on spoken documents with transcription errors [1], contrary to more sophisticated information about syntactic structures of sentences.

One reason that explains that POS information is not usually very helpful for transcription is related to the combination of POS knowledge at the LM level. The interpolation between a word level N-gram LM and a POS level N-class model has demonstrated a limited effectiveness [2, 3]. In [4], a 3-gram LM is defined over word/tag pairs rather than words and the recognition problem is defined as finding the best joint word and POS tag sequences. This approach results in a significant word error rate (WER) reduction. However, due to the drastic increase of entries in the LM, the approach requires a very large amount of training data and heavily relies on smoothing techniques to make up for the lack of data.

We propose a slightly different approach where POS information is combined with the LM score in a post-processing stage rather than integrated in the LM as in previous approaches. An interesting point is that our method does not require a large amount of annotated training data as opposed to [4]. The idea is to process N-best lists of sentence hypotheses. We first disambiguate each hypothesis; a score function combining the POS, LM, and acoustic scores is then used to generate a transcription from the N-best lists.

In this paper, we first describe the proposed combined score function. We then present our baseline system and the morphosyntactic tagger used in the experiments. Experimental results demonstrate that a significant improvement of the transcription can be obtained using our combined score function with various decoding criteria. We finally investigate the use of morphosyntactic information for confidence measures.

2. Combined score function

Throughout this study, morphosyntactic information is used in a post-processing stage to process N-best sentence hypothesis lists. Although N-best lists are not as informative as word graphs, each entry can be seen as a standard text, permitting thus POS tagging. Besides, processing word graphs would require a huge increase of computation as each word must be extended by all the possible POS tags.

To combine morphosyntactic information with the LM and acoustic scores, we use a tagger to determine the most likely POS tag sequence t_1^m corresponding to a sentence hypothesis w_1^n. Based on this information, we compute the morphosyntactic probability of the sentence hypothesis

$$P(t_1^m) \approx \prod_{i=1}^{m} P(t_i | t_{i-N+1}^{i-1}) \ . \tag{1}$$

Note that the number m of tags may differ from the number n of words as we associate a unique POS with locutions, consecutive proper names, or cardinals.

We extend the global score of a sentence by adding the morphosyntactic score to the score computed in practice by most ASR systems with an appropriate weight. The combined score for a sentence w_1^n, corresponding to the acoustic input y_1^t, is therefore given by

$$\begin{aligned} s(w_1^n) \quad = \quad & \log P(y_1^t | w_1^n) + \alpha \log P(w_1^n) \\ & + \beta \log P(t_1^m) + \gamma n \end{aligned} \tag{2}$$

where α is the LM scale factor, β is the POS scale factor and γ is the word insertion penalty. $P(y_1^t | w_1^n)$ is the acoustic score given by the ASR system and $P(w_1^n)$ is the LM probability as given by a N-gram language model. Introducing POS information at the sentence level allows us to differently tokenize

sequences of words and tags and to more explicitly penalize unlikely sequences of tags like a plural noun following a singular adjective.

Based on the score function defined in (2), which includes all the available sources of knowledge, we can decode N-best lists using various criteria. We considered three criteria, namely maximum a posteriori (MAP), minimum expected word error rate [5], and consensus decoding on N-best lists [6]. The two last criteria, often used in current systems, aim at reducing the word error rate but increase the sentence error rate. Results are presented in section 4. The posterior sentence probabilities used in the two word error minimization criteria are also extensively used to derive confidence measures from N-best lists or word graphs [7]. We study in section 5 the impact of morphosyntactic information on confidence measures.

3. Experimental setup

Experiments are carried out on the Ester corpus, which consists of French radio broadcast news [8]. A development set of 4 hours from 4 different broadcasters was used to empirically determine the optimal values for the scale factors α, β and the insertion penalty γ. Tests are carried out on a separate set of 4 hours from the same broadcasters. Broadcast news shows were manually segmented into breath groups to avoid problems due to poor morphosyntactic segmentation. However, automatic segmentation based on filler models in a phone-loop decoder provides results comparable to the manual segmentation and should therefore not modify the conclusions.

The N best sentence candidates were generated based on our 64 k word broadcast news transcription system according to the following procedure. A large initial word graph is first generated using context-independent phone models along with a 3-gram LM. The word graph is then expanded with a 4-gram LM and rescored with context-dependent phone models. A last pass generates a final word graph with adapted triphone models. N-best sentence hypothesis lists are generated from the final word graph by extracting the N best sentence candidates, where each sentence differs with at least one word from the other ones. The language model probabilities were estimated on 350 M words from the French newspaper "Le Monde" and interpolated with LM probabilities estimated over 1 M words corresponding to the reference transcription of 80 hours of training material[1].

For decoding purposes, lists of 100 sentence hypotheses were generated. To speed up computation, the 4-gram expansion of the initial word graph was limited to the extraction of the 1000 best paths. For confidence measures, 1000 sentence candidates were considered and no limitation was applied in the initial word graph expansion.

POS tagging was carried out with a statistical tagger adapted to French broadcast news, based on a 3-gram class model whose probabilities were estimated over a corpus of 200 k words of reference broadcast news transcriptions. This tagger was evaluated on manual and automatic transcriptions and exhibits a correct tag rate of 95.7 % on both[2], a recognition rate comparable to those obtained on written documents. The initial tag set contains 95 tags corresponding to grammatical classes, along with information about gender and number. The tag set was extended by adding specific classes for the 100 most frequent function words, as preliminary experiments showed a better improvement of WER with this tag set. Finally, morphosyntactic scores are obtained from a 7-gram class model. The reason for choosing a model with longer dependencies than the model used for tagging is to capture longer syntactic contexts than with the word-based language model.

[1]The authors would like to thank François Yvon for providing the LM and the lexicon.

[2]On automatic transcriptions, the tag rate is computed only for the words that are correctly recognized.

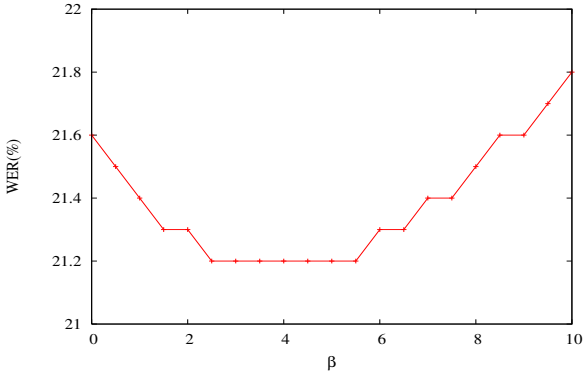

Figure 1: *WER as a function of β on development data.*

baseline ASR system	19.9
contextual prob. + disambig.	19.1
lexical and contextual prob. + disambig.	19.0
lexical and contextual prob.	19.0
class-based LM	19.5

Table 1: *WER(%) on test data.*

4. Decoding with N-best lists

We first compare our strategy for reordering N-best lists with standard MAP approaches before studying WER minimization. We show that the observed improvement is significant and compare our method to an approach based on homophone extensions.

4.1. MAP criterion

For each breath group, we select the best hypothesis $w^{(i)}$ that maximizes $s(w^{(i)})$ as given by (2). Figure 1 shows the WER on the development data as a function of the POS scale factor β, the other parameters being optimized separately for each value of β. Results clearly show an improvement when morphosyntactic information is taken into account. This result is confirmed on the test corpus where our approach achieves an absolute decrease of 0.8 % of the WER as reported in Tab. 1, lines 1 and 2.

We compared our approach with class-based LMs incorporated in the transcription process by linear interpolation with a word-based LM according to

$$P(w_1^n) = \prod_{i=1}^{n} [\lambda P_{\text{word}}(w_i|w_1^{i-1}) + (1-\lambda) P_{\text{POS}}(w_i|w_1^{i-1})]$$

with

$$P_{\text{POS}}(w_i|w_1^{i-1}) = \sum_{t_{i-N+1}\ldots t_i} P(w_i|t_i) P(t_i|t_{i-N+1}^{i-1}) \quad . \quad (3)$$

We reranked the N-best lists by interpolating the N-class based POS tagger and the word level language model, the interpolation factor λ being optimized on the development data. We

noticed an absolute decrease of 0.4 % w.r.t. the baseline system, *i.e.*, half of the decrease previously observed (Tab. 1, last line).

One of the differences between the two approaches is the use of the lexical probabilities $P(w_i|t_i)$ in (3). To study the influence of the lexical probabilities, two new score functions were defined. The first one, defined as

$$s'(w_1^n) = \log P(y_1^t|w_1^n) + \alpha \log P(w_1^n) + \beta[\sum_{i=1}^{n} \log P(w_i|t_i) + \log P(t_1^m)] + \gamma n \quad (4)$$

takes into account the lexical probabilities after disambiguation. The second one considers all the possible sequences of tags rather than the best one. It is defined as

$$s''(w_1^n) = \log P(y_1^t|w_1^n) + \alpha \log P(w_1^n) + \beta[\sum_{i=1}^{n} \log P_{\text{POS}}(w_i|w_1^{i-1})] + \gamma n \quad (5)$$

and simply corresponds to a linear interpolation of the log probabilities of the word-based and POS-based LMs. Results reported in Tab. 1 (resp. lines 3 and 4) show a slight improvement of the WER by taking into account contextual probabilities. Although they do not demonstrate the interest of disambiguation, they clearly establish that linear interpolation of log probabilities is more effective than that of probabilities.

4.2. Word error minimization criteria

Combined scores incorporating morphosyntactic information can be used to decode N-best lists using decoding criteria that aim at minimizing the word error rate, rather than the sentence error rate as the MAP criterion does. Two popular criteria can be used to explicitly minimize word error rates: the first one consists in approximating the posterior expectation of the word error rate by comparing each pair of hypotheses in the N-best list [5]; the second one, consensus decoding, is based on the multiple alignment of the N-best hypotheses into a confusion network [6].

Both criteria rely on the computation of the posterior probability for each sentence hypothesis $w^{(i)}$

$$P(w^{(i)}|y_1^t) = \frac{e^{s(w^{(i)})/z}}{\sum_j e^{s(w^{(j)})/z}} \quad (6)$$

where the posterior probability can be computed from a score including morphosyntactic knowledge. The combined score is scaled by a factor z in order to avoid over-peaked posterior probabilities. In the experiments described below, the combined score with lexical probabilities defined by (4) was used.

Results are reported in Tab. 2 for the three decoding criteria, namely MAP, WER minimization, and consensus, with and without POS knowledge. In both cases, we observe a slight decrease of the WER when using word error minimization criteria, along with an increased sentence error rate. However, the gain observed is not meaningful because of the limited size of the N-best lists (N=100); a larger gain was observed with a 1000-best list, as illustrated in Tab 3. Interestingly, results indicate that including morphosyntactic information also benefits to word error minimization criteria, thus indicating that the WER improvement due to POS tags is consistent whatever the decoding criterion used.

4.3. Discussion on the results

Statistical tests were carried out to measure the significance of the WER improvement observed, assuming independence of the errors across breath groups. Both the paired t-test and the paired Wilcoxon test resulted in a confidence over 99.9 % for all the decoding criteria w.r.t. the baseline system. Besides, for the MAP criterion, the same tests indicate that global scores computed as (2), (4), or (5) led to a significant improvement w.r.t. interpolated class-based LM with a confidence over 99 %.

The robustness of morphosyntactic tagging may be questionable for spontaneous speech. To evaluate the interest of our method on more spontaneous speech, performance was measured on a short extract of 3,650 words containing interviews with numerous disfluencies. The baseline WER of 46.3 % is reduced to 44.5 % with (2) and to 44.3 % with (4) using the MAP criterion. This 4% relative improvement is consistent with the relative improvement obtained on the entire test set, thus demonstrating that the method is robust to speaking style.

To conclude this section, we observed that changes induced by POS knowledge generally produce more grammatical utterance transcriptions as indicated by the sentence error rates reported in Tab. 2. Indeed, an analysis of the sentence error rate shows a significant reduction when taking into account morphosyntactic information. In particular, we noticed several corrections of agreement or tense errors such as "*une date qui À DONNER le vertige à une partie de la France*" ("*a date which TO GIVE a part of France fever*").

4.4. Comparison with an approach on homophones

We emphasized in the introduction the interest of morphosyntactic information to correct gender and number agreement for homophones. Based on this principle, an alternative data representation, lattices of homophones, has been previously suggested to take into account POS knowledge [9]. We compared the pertinence of these lattices by extending the best hypothesis found by our baseline ASR system by the homophones of its words. As the number of possible homophones of short words can be huge, we only considered adjectives, common nouns, verbs, and personal pronouns, and limited their extensions to homophones associated with the same lemma. For instance, "*un texte qui finis*" ("*a text that ends*") is extended by "*un textes qui finis*", "*un textes qui finit*", and "*un texte qui finit*". By using global score (4) according to a MAP criterion, we noticed a reduction of WER from 19.9 % to 19.6 %, which is not as high as the one previously obtained from N-best lists.

5. Confidence measures

In the previous experiments, we quantitatively demonstrated the interest of morphosyntactic information to significantly improve transcription by post-processing of N-best lists. In particular, we showed that POS information is effective to compute sentence posterior probabilities for N-best lists decoding with WER minimization criteria. As sentence posterior probabilities are commonly used to derive confidence measures from N-best lists or lattices, we study in this section the impact of morphosyntactic information on confidence measures.

When consensus decoding is used, confidence measures are directly taken as the posterior probabilities of the best word for each slot in the confusion network. When MAP decoding is used, confidence measures are computed from the sentence posterior probabilities as in [7]. The idea is to align each alternate sentence hypothesis with the best one. The confidence measure

	WER			SER		
	MAP	min. WE	cons.	MAP	min. WE	cons.
without POS	19.9	19.8	19.8	61.8	62.2	62.4
with POS	19.0	18.9	18.9	59.4	59.6	59.7

Table 2: *WER(%) and sentence error rate (%) on test data for various decoding techniques.*

	MAP decoding		consensus decoding	
	w/o POS	w/ POS	w/o POS	w/ POS
WER	19.7	18.7	19.4	18.6
NCE w/o POS	0.307	0.265	0.198	–
NCE w/ POS	0.326	0.288	–	0.211

Table 3: *WER(%) and normalized cross entropy for various decoding techniques with and without POS score.*

for a word $w_i = w$ in the best sentence hypothesis is the sum of the sentence posterior probabilities over all the sentences that contain the same word w aligned with w_i.

Confidence measures are computed from 1000-best lists using the combined score defined in (4). In the case of MAP decoding, the scaling factors and insertion penalty used for the computation of the sentence posteriors are different from those used for reordering the N-best lists and were optimized on the development set to maximize the normalized cross entropy (NCE). In MAP decoding, it is therefore possible to use morphosyntactic information only for reordering or only for the purpose of confidence measure computation.

Tab. 3 summarizes the results in terms of NCE where the NCE reflects the quality of the confidence measures w.r.t. an optimal constant confidence measures. The word error rates for the various configuration tested, namely MAP/consensus decoding with/without POS information, are reported on the first line. The next two lines report NCE obtained when computing confidence measures respectively without and with morphosyntactic information. These results clearly demonstrate that morphosyntactic information improves confidence measures, as it provides additional syntactic information not sufficiently captured by the word based LM. Indeed, we observed on several examples that a plot of the local score $P(t_i|t_{i-N+1}^{i-1})$ exhibits a significant decrease on erroneous words whereas the language model may exhibit the same behavior on correct words due to smoothing and back-off. The detection error trade-off curves also reveal the increased quality of the confidence measures with morphosyntactic information, in particular at low miss detection (correct word removed) rates. Finally, one can note that the quality of the confidence measure is lower for consensus decoding than for MAP decoding. This is due to the fact that the parameters, and in particular the scale factor z in (6), were optimized to minimize the word error rate in the first case while they were optimized to maximize the NCE in the second case, since different parameters can be used for rescoring and for confidence measure computation in MAP decoding.

6. Conclusions

We presented a study on the integration of morphosyntactic knowledge at the sentence level. As opposed to previous studies that introduced morphosyntactic knowledge at the LM level by interpolation, we combine at the sentence level the acoustic, LM, and morphosyntactic log scores. The resulting score function was found to significantly decrease the WER both with maximum a posteriori and explicit WER minimization decoding criteria. Sentence posterior probabilities incorporating morphosyntactic knowledge also resulted in improved confidence measures.

In addition to the fact that morphosyntactic processing of N-best lists provides syntactic information on top of the lexical transcription, we observed that the latter is generally more grammatical as it exhibits lower sentence error rates. This result is particularly interesting since more grammatical transcriptions should make easier exploiting the transcription with natural language processing techniques for high level semantic analysis.

Finally, results not included in the scope of this paper on the use of morphosyntactic information for the English language in an handwriting recognition experiments demonstrate the interest of the proposed technique for less inflectional languages, even though the relative improvement is unsurprisingly lower for the English language than for the French one.

7. References

[1] S. Huet, G. Gravier, and P. Sébillot, "Are morphosyntactic taggers suitable to improve automatic transcription?" in *Proc. of TSD*, 2006.

[2] F. Jelinek, *Readings in Speech Recognition*. Morgan Kaufmann Publishers, 1990, ch. Self-Organized Language Modeling for Speech Recognition, pp. 450–506.

[3] G. Maltese and F. Mancini, "An automatic technique to include grammatical and morphological information in a trigram-based statistical language model," in *Proc. of ICASSP*, 1992.

[4] P. A. Heeman, "POS tags and decision trees for language modeling," in *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[5] A. Stolcke, Y. König, and M. Weintraub, "Explicit word error minimization in N-best list rescoring," in *Proc. of Europeech*, 1997.

[6] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[7] B. Rueber, "Obtaining confidence measures from sentence probabilities," in *Proc. of Eurospeech*, 1997.

[8] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news," in *Proc. of Eurospeech*, 2005.

[9] J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk, "Where are we in transcribing French broadcast news?" in *Proc. of Eurospeech*, 2005.