Toward the Integration of Natural Language Processing and Automatic Speech Recognition: Using Morpho-Syntax and Pragmatics for Transcription

Stéphane Huet, Gwénolé Lecorvé, Guillaume Gravier, Pascale Sébillot

1

Toward the Integration of Natural Language Processing and Automatic Speech Recognition: Using Morpho-Syntax and Pragmatics for Transcription

Stéphane Huet[1], Gwénolé Lecorvé[2], Guillaume Gravier[3], and Pascale Sébillot[2]

[1] IRISA, Université de Rennes 1
[2] IRISA, INSA
[3] IRISA, CNRS

In the framework of multimedia analysis and interaction, speech and language processing plays a major role. Many multimedia documents contain speech from which high level semantic information can be extracted, as in broadcast news or sports videos, with typical applications such as spoken document indexing, topic tracking and summarization. Hence, many multimedia document analysis applications require a collaboration between speech recognition and natural language processing (NLP) techniques. As NLP techniques are traditionally designed for text analysis, this combination can be seen as a multimodal fusion issue where the two modalities are audio and text. However, most of the time, both modalities are considered sequentially. A typical approach consists in automatically transcribing the audio track before analyzing the output—here considered as a regular text—with NLP methods. Independently processing the two modalities clearly seems suboptimal. This chapter focuses on recent research work toward a better integration between automatic speech recognition (ASR) and NLP for the analysis of spoken multimedia documents with the goal of achieving a better transcription of multimedia streams.

The speech processing and text processing communities have had a long history of misunderstanding, mostly due to two different approaches to natural language: a pure statistical one vs. a more symbolic, rule-based one. But the last 15 years have begun to re-appropriate the joint use of ASR and NLP. If using ASR and NLP is now a clear will, the cooperation is not that simple. First, oral output has characteristics, such as repetitions, revisions or fillers, known as disfluencies, that make it difficult. Moreover, additional difficulties come from the fact that automatic transcriptions are not segmented into sentences—the equivalents of shots for texts—, lack punctuation and, in

the case of some ASR systems, capitalization. Finally, transcription errors might impact text processing techniques.

The problem of combining the ASR and NLP can be tackled in several ways. A popular one consists in reformatting the automatic transcription to look like a regular text using re-punctuation techniques and correcting disfluencies [15]. A second possibility is to adapt NLP techniques to take into account additional information provided by the ASR system, such as word-level confidence measures or word graphs [2]. We believe that these approaches cannot replace a better and deeper integration between ASR and NLP: for example, re-punctuation cannot help NLP recover from transcription errors. This chapter proposes a reflection and research tracks toward this goal, considering the use in ASR of linguistic knowledge that are mostly absent from current transcription systems.

Different kinds of linguistic knowledge have been considered for integration into ASR systems, namely morphological, syntactic, semantic and pragmatic, as reviewed in Section 1.2. However, most approaches consider minor changes of the ASR system (e.g., by modifying the language model) rather than a real in-depth integration. We explore in this chapter two instances of a better combination of ASR and NLP, considering morpho-syntactic information in Section 1.3 and pragmatic information for unsupervised language model adaptation in Section 1.4. Clearly, the main idea underlying this work is to take into account multimodal cues at each step of the spoken document analysis process, for example to correct transcription errors using NLP knowledge, to segment multimedia streams into topics (see Section 1.4.2) or to adapt the ASR system to the current topic.

1.1 The basic principles of automatic speech recognition

Before considering the use of linguistic information in ASR systems, we review the fundamentals of speech recognition and briefly describe the experimental framework common to the two experiments described in Sections 1.3 and 1.4.

1.1.1 General principles

Most automatic speech recognition systems rely on statistical models of speech and language to find out the best transcription, i.e., word sequence, given a (representation of the) signal y, according to

$$\widehat{w} = \arg\max_{w} p(y|w)\ P[w]\ \ . \tag{1.1}$$

Language models (LM), briefly described below, are used to get the prior probability $P[w]$ of a word sequence w. Acoustic models, typically continuous density hidden Markov models (HMM) representing phones, are used to compute the probability of the acoustic material for a given word sequence,

$p(y|w)$. The relation between words and acoustic models of phone-like units is provided by a pronunciation dictionary which lists the words known to the ASR system along with the corresponding pronunciations. Hence, ASR systems operate on a closed vocabulary whose typical size is between 60,000 and 100,000 words or tokens. Because of the limited size of the vocabulary, word normalization, by ignoring the case for example or by breaking compound words, is often used to limit the number of out-of-vocabulary words. The consequence is that the vocabulary of an ASR system is not necessarily suited for natural language processing.

As mentioned previously, the role of the language model is to define a probability distribution over the set of possible sentences according to the vocabulary of the system. As such, the language model is a key component for a better integration between ASR and NLP. ASR systems typically rely on N-gram based language models because of their simplicity which makes the maximization in (1.1) tractable. The N-gram model defines the probability of a sentence w_1^n as

$$P[w_1^n] = \prod_{i=1}^{n} P[w_i|w_{i-N+1}^{i-1}] \ , \tag{1.2}$$

where the probabilities of the sequences of N words $P[w_i|w_{i-N+1}^{i-1}]$ are estimated from large text corpora. Because of the large size of the vocabulary, observing all the possible sequences of N words is impossible. A first approach to circumvent the problem is based on smoothing techniques, such as discounting and back-off, to avoid null probabilities for events unobserved in the training corpus. Another approach rely on N-gram models based on classes of words [5] where a N-gram model operates on a limited set of classes, and words belong to one or several classes. The probability of a word sequence is then given by

$$P[w_1^n] = \sum_{t_1 \in \mathcal{C}(w_1) \ldots t_n \in \mathcal{C}(w_n)} \prod_{i=1}^{n} P[w_i|t_i] P[t_i|t_{i-N+1}^{i-1}] \ , \tag{1.3}$$

where $\mathcal{C}(w)$ denotes the set of possible classes for a word w.

In practice, (1.1) is evaluated in the log-domain and the LM probabilities are scaled in order to be comparable to acoustic likelihoods, thus resulting in the following maximization problem

$$\widehat{w} = \arg\max_{w} \ \ln p(y|w) + \beta \ \ln P[w] + \gamma \ |w| \ , \tag{1.4}$$

where the LM scale factor β and the word insertion penalty γ are empirically set.

The ultimate output of an ASR system is obviously the transcription. However, additional information, such as confidence measures or transcription alternatives, can also be obtained. This information might prove useful for NLP as it can help to avoid error-prone hard decisions from the ASR system.

Rather than finding out the best word sequence maximizing (1.4), one can output a list of the $\mathcal{N}$-best word sequences thus keeping track of the alternative transcriptions that were discarded by the system. For a very large number of transcription hypotheses, these $\mathcal{N}$-best lists can be conveniently organized as word graphs where each arc corresponds to a word. From the set of alternative hypotheses, confidence measures can be computed for each word, where the measures reflect how confident is the system.

1.1.2 The IRENE broadcast news transcription system

The IRENE broadcast news transcription system, jointly developed by IRISA and ENST for the ESTER broadcast news transcription evaluation campaign [8], implements the basic principles described in the previous section after a partitioning step which aims at segmenting the input stream into pseudo-sentences. The system has a vocabulary of 64,000 words.

Regions containing speech are first detected before performing a further partitioning into speaker turns. Since (1.4) can only be solved for short utterances, the speech stream is finally segmented into breath-groups based on the energy profile in order to detect breath intakes[4]. Let us stress the fact this segmentation is not based on syntactic and grammatical considerations, even though breath pauses and grammar are related.

Transcription itself is carried out in three passes. A first pass with fairly simple context-independent acoustic models and a 3-gram word based LM aims at generating large word graphs. These word graphs are then rescored with more complex context-dependent acoustic models and a 4-gram LM. Rescoring word graphs is based on (1.4) where the maximization is limited to the set of word sequences encoded in the word graph, thus making the use of more complex models tractable. Finally, based on the transcription from the second pass and the speaker partition obtained in the segmentation step, the acoustic models are adapted for each speaker and final word graphs are obtained by rescoring the initial word graphs with speaker-adapted acoustic models.

Experiments reported in this chapter were carried out on the ESTER French broadcast news transcription task. A corpus of about one hundred hours of manually transcribed data was used and divided into three parts: a large part was reserved for the purpose of acoustic and language model training while two sets of four hours each, from four different broadcasters, were used as development and test sets respectively. The development set was used to tune the many parameters of the ASR system such as the language model scale factor. The language model was obtained by interpolating a LM estimated on 1 million words from the manual transcriptions of the training set with a LM estimated from 350 million words from the French newspaper *Le Monde*.

[4] To avoid problems due to segmentation errors, the entire partitioning process was done manually in the experiments reported in this chapter.

1.2 Fusion of text and speech modalities: an overview

Let us come back to the problem of combining the text and audio modalities. We review in this section the literature concerning the use of linguistic knowledge in ASR systems successively considering morphology, syntax, semantics and pragmatics.

Morphology considers the structure of words. Morphological analyzers can be used to convert words into their canonical form, e.g., a lemma or a stem. Such knowledge is incorporated in ASR systems by defining a LM over canonical forms rather than words, which is convenient in order to reduce the vocabulary size in particular for agglutinative or morphologically rich languages. Factored models[5] have been specifically developed to integrate morphological components as factors in the language model probability computation, where the factors can be stems, morphological classes or even the words themselves [32].

Syntax considers the structure of sentences and syntagms, e.g., nominal or verbal groups. A first possibility relies on part-of-speech (POS) information, i.e., grammatical classes such as noun, verb, and preposition, associated with each word, known as POS tags. A class-based LM can be defined over POS tags and combined with a word-based LM [16]. The main interest of POS-based LMs is the limited number of tags with respect to the number of words and their ability to point out ungrammatical word (actually tag) sequences. Moreover, morphological knowledge can also be included in the tags leading to morpho-syntactic information. A second use of syntactic information is to extract locutions based on the statistical study of co-occurrences [25] or the use of regular expressions [20]. Such locutions are included in the vocabulary of the ASR system as multi-word units. Finally, syntactic analysis of transcription hypotheses can also be done in order to choose the most grammatical ones. As designing generic syntactic parsers robust to transcription errors is an awfully difficult task, systems either complex [6] or limited to a specific domain [22] have been proposed.

Semantics considers the meaning of the words and the relations between words. Few works include semantic information in ASR systems but relations between words can be incorporated in long-span language models as in [27] and [3]. The idea is to put forward sentence hypotheses containing words with related meanings. Relations between words are automatically acquired either considering co-occurrences in syntagms or text windows, or considering words sharing the same neighbors. However, long-span language models are difficult to integrate in an ASR system.

Finally, pragmatics considers the context, shared by the redactor and the reader, so that the document makes sense. The topic of the document is a typical example of pragmatic knowledge which can be used in ASR systems,

[5] Factored model are similar to factorial Markov models where the state space is distributed over a set of factors.

for instance for LM and vocabulary adaptation. One approach for LM topic adaptation relies on a set of predefined domain-specific LMs [13, 9]. However, this method requires the *a priori* definition of the set of possible topics. Another solution is to gather a specific adaptation corpus for each document, either by selecting a subset of a very general corpus [14] or by collecting texts on the Internet [26].

Whatever the type of knowledge, most techniques naturally rely on an integration at the language model level. A typical approach consists in modifying the word-based N-gram LM, for example using interpolation techniques. However, this approach implies only minor modifications of the architecture of the ASR system and thus often only yield marginal improvements.

In this chapter, we report on work targeting a better integration of the text and speech modalities for two different sources of knowledge, namely morpho-syntax and pragmatics, where topic adaptation is considered in the latter. These two types of linguistic information are crucial for multimedia applications. Morpho-syntactic knowledge enables more grammatical transcriptions, thus facilitating the use of *a posteriori* NLP techniques on the output. Topic adaptation is vital for the accurate transcription of multimedia streams where various topics can be found.

1.3 Morpho-syntactic knowledge integration

In this section, we present our method to integrate morpho-syntactic information in the ASR process. As mentioned in the previous section, part-of-speech tags along with morphological knowledge about gender, number, tense, mode or case are used to convey morpho-syntactic information. Previous work combining class-based LMs and word-based ones have demonstrated a limited effectiveness [33]. In [10], a 3-gram LM is built over word/tag pairs rather than words and the recognition problem is redefined as finding the best joint word and POS tag sequences. This approach results in a significant reduction of the word error rate (WER) but requires very large amount of training data for the LM and heavily relies on smoothing techniques.

We propose a different approach where POS information is combined with the LM score in a post-processing stage of a $\mathcal{N}$-best list of hypotheses rather than integrated in the LM as in previous approaches. The basic idea is to tag the output of the ASR system in order to favor the hypotheses with correct POS sequences, like a singular noun following a singular adjective. Closely related to [10], our method does however not require a large amount of annotated training data. In this section, we demonstrate that POS tagging can be reliably applied to automatic transcriptions and that the resulting tags can actually improve the word error rate and confidence measures.

1.3.1 Morpho-syntactic tagging of automatic transcriptions

Morpho-syntactic tagging is a widely used technique in NLP and taggers are now considered as reliable enough to automatically tag a text according to POS information. However, most experiments were carried out on written text, and spoken corpora on the contrary have been seldom studied. As oral output has specificities that are likely to disturb taggers, we first demonstrate that such noisy texts can be reliably tagged.

We built a morpho-syntactic tagger based on the popular technique of HMM [19], where tagging is expressed as finding out, for each sentence, the most probable POS tag sequence, among all the possible sequences according to a lexicon. In order to adapt the model to the characteristics of oral, we used a 200,000-word training set from the manual transcriptions of the training corpus. Moreover, we removed all capital letters and punctuation marks to obtain a format similar to a transcription and segmented the set into breath-groups. We also restrained the vocabulary of the tagger to the one of our ASR system. We chose our POS tags in order 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 93 tags.

To quantitatively evaluate morpho-syntactic tagging, we manually tagged a one hour broadcast. We first investigated the behavior of the tagger on manually transcribed text by comparing the tag found for each word with the one of the reference. For automatic transcriptions, evaluating the tagger is more problematic than for manual transcriptions since ASR output contains misrecognized words; for the hypotheses containing grammatical errors, it becomes impossible to know which sequence of POS would be right. We therefore compute the tag rate only for the words that are correctly recognized.

Table 1.1, first line, reports results obtained on the one hour corpus with our tagger, where the WER on the transcription is 22.0 %. We achieved a tag accuracy over 95 % which is comparable to the results usually given on written corpora. Furthermore, similar performance level are obtained on both the manual and automatic transcriptions, which establishes therefore that morpho-syntactic tagging is reliable, even for text produced by an ASR system whose recognition errors are likely to jeopardize the tagging of correctly recognized words. The robustness of tagging is explained by the fact that tags are locally assigned. We compared the performances of our tagger with those of Cordial[6], one of the best taggers available for written French and which has already produced good results on a spoken corpus [30]. Results reported in the last line of Table 1.1 are comparable with our HMM-based tagger when we ignore confusion between proper names and common names. Indeed, the lack of capital letters is particularly problematic for Cordial, which relies on this information to detect proper names.

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

transcription	manual	automatic
HMM tagger	95.7 (95.9)	95.7 (95.9)
Cordial	90.7 (95.0)	90.6 (95.2)

Table 1.1. Tag accuracy (in %), where results between parentheses are computed when confusion between common names and proper names is ignored.

1.3.2 Reranking of $\mathcal{N}$-best lists

Morpho-syntactic information is here used to post-processing $\mathcal{N}$-best sentence hypothesis lists. Although $\mathcal{N}$-best lists are not as informative as word graphs, each entry can be seen as a standard text, permitting thus POS tagging.

To combine morpho-syntactic information with the LM and acoustic scores, we first 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 morpho-syntactic probability of the sentence hypothesis

$$P[t_1^m] = \prod_{i=1}^{m} P[t_i | t_{i-N+1}^{i-1}] \ . \tag{1.5}$$

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. To take into account longer dependencies than the 4-gram word-based LM, we chose a 7-gram POS-based LM.

We propose a new global score of a sentence [12] by adding the morpho-syntactic score to the score given in (1.4) 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

$$s(w_1^n) = \log p(y_1^t | w_1^n) + \alpha \log P[w_1^n] + \beta \log P[t_1^m] + \gamma n \ . \tag{1.6}$$

Integrating 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 (1.6), which includes all the available sources of knowledge, we can reorder $\mathcal{N}$-best lists using various criteria. We considered three criteria, namely maximum a posteriori (MAP), minimum expected word error rate [24] and consensus decoding on $\mathcal{N}$-best lists [17]. The two last criteria, often used in current systems, aim at reducing the word error rate at the expense of an increased sentence error rate (SER).

MAP criterion

The MAP criterion selects among the $\mathcal{N}$-best list generated for each breath-group the best hypothesis $w^{(i)}$ which maximizes $s(w^{(i)})$ as given by (1.6).

baseline ASR system	contextual probabilities	lexical and contextual probabilities	class-based LM
19.9	19.1	19.0	19.5

Table 1.2. WER (in %) on test data obtained with a LM limited to a word-based LM (1st column) or with an ASR system including POS according to equations (1.6), (1.7) or (1.8) (last three columns).

Results on the test corpus show that our approach achieves an absolute decrease of 0.8 % of the WER as reported in Tab. 1.2, columns 1 and 2. By taking into account lexical probabilities $P[w_i|t_i]$, which are usually included in class-based LM, we observed a minor additional decrease (Tab. 1.2, column 3) of the WER. The score in this last case is computed by linearly interpolating log-probabilities by

$$\begin{aligned} s'(w_1^n) = \log P(y_1^t|w_1^n) + \alpha \log P[w_1^n] \\ +\beta \left(\sum_{i=1}^{n} \log P[w_i|t_i] + \log P[t_1^m] \right) + \gamma n \end{aligned} \tag{1.7}$$

and tends to penalize words that are rarely associated with the proposed tag.

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

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

with

$$P_{\text{pos}}[w_i|w_1^{i-1}] = \sum_{t_{i-N+1} \ldots t_i} P[w_i|t_i] \;\; P[t_i|t_{i-N+1}^{i-1}] \;\; . \tag{1.8}$$

We reevaluated the $\mathcal{N}$-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 % with respect to the baseline system, i.e., half of the decrease previously observed (Tab. 1.2, last column). The better improvement of WER with our method clearly establishes that linear interpolation of log probabilities is more effective than that of probabilities.

Word error minimization criteria

Combined scores incorporating morpho-syntactic information can be used to reorder $\mathcal{N}$-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 the WER: the first one consists in approximating the posterior expectation of the word error

	WER			SER		
	MAP dec.	min. WE	cons. dec.	MAP dec.	min. WE	cons. dec.
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 1.3. Word (WER) and sentence (SER) error rates (in %) on the test data for various decoding techniques.

rate by comparing each pair of hypotheses in the $\mathcal{N}$-best list [24]; the second one, consensus decoding, is based on the multiple alignment of the $\mathcal{N}$-best hypotheses into a confusion network [17].

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

$$P[w^{(i)}|y_1^t] = \frac{e^{s(w^{(i)})/z}}{\sum_j e^{s(w^{(j)})/z}} \tag{1.9}$$

where the posterior probability is obtained from a score including morpho-syntactic knowledge, the one given by Eq. (1.7) in our case. The combined score is scaled by a factor z in order to avoid over-peaked posterior probabilities.

Results are reported in Tab. 1.3 for the three decoding criteria, namely MAP, WER minimization and consensus, with and without POS knowledge. In both cases, we observe a slight WER improvement when using word error minimization criteria, along with an increased SER. However, the gain observed is marginal because of the limited size of the $\mathcal{N}$-best lists ($\mathcal{N}$=100). Indeed, with $\mathcal{N} = 1000$ the WER decreased from 19.7 % to 19.4 % without POS. A more limited gain was observed when using POS with a decrease from 18.7 % to 18.6 %.

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. For all the decoding criteria, both the paired t-test and the paired Wilcoxon test resulted in a confidence over 99.9 % that the difference of WER by using or not POS knowledge is not observed by chance. Besides, for the MAP criterion, the same tests indicate that global scores computed as (1.6) or (1.7) led to a significant improvement with respect to the interpolated class-based LM with a confidence over 99 %.

We observed that our method is robust for spontaneous speech. Indeed, we measured performance on a short extract of 3,650 words containing interviews with numerous disfluencies and observed that the baseline WER of 46.3 % is reduced to 44.5 % with (1.6) and to 44.3 % with (1.7) using the

	WER	NCE without POS	NCE with POS
decoding without POS	19.7	0.307	0.326
decoding with POS	18.7	0.265	0.288

Table 1.4. WER (in %) and normalized cross entropy for MAP decoding with and without POS score.

MAP criterion. This 4% relative improvement is consistent with the relative improvement obtained on the entire test set. Additional experiments with automatic segmentation also demonstrated the validity of our approach in that case.

Experiments reported here were carried out on the French language, whose nouns, adjectives and verbs are very often inflected for number, gender or tense into various homophone forms. However, experiments conducted to improve a hand-writing recognition system in the English language, show that morpho-syntactic knowledge still brings an WER improvement, even though English is less inflected than French.

To conclude this section, we observed that introducing morpho-syntactic knowledge in the ASR system yield more grammatically correct utterance transcriptions as indicated by the SERs reported in Tab. 1.3. 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*").

1.3.3 Confidence measures

We have shown how POS knowledge can reduce transcript errors. Another interest of morpho-syntactic information for ASR systems is that it can bring new information to compute confidence measures.

Plots of the conditional probabilities $P[t_i|t_{i-N+1}^{i-1}]$ for POS sequences and $P[w_i|w_{i-M+1}^{i-1}]$ for word sequences show that $P[t_i|t_{i-N+1}^{i-1}]$ exhibits a significant decrease on erroneous words where language model may show the same behavior on correct words due to smoothing or back-off. This property is particularly interesting to compute confidence measures. As sentence posterior probabilities are commonly used to derive confidence measures from $\mathcal{N}$-best lists or lattices, we compute them as in [21].

Confidence measures are obtained from $1,000$-best lists using the combined score (1.7). We limit the study to the lists obtained with MAP decoding for which the lowest SER was achieved. The scaling factors and insertion penalty used for the computation of the sentence posteriors are different from those used for reordering the $\mathcal{N}$-best lists and were optimized on the development set to maximize the normalized cross entropy (NCE), a commonly used indicator to evaluate confidence measure on the correctness of a word.

Table 1.4 summarizes the results, where the higher the NCE, the better the confidence measure. WER with and without POS information are given

in the first column. The next two columns report NCE obtained when computing confidence measures respectively without and with morpho-syntactic information. Results show that POS improves confidence measures in both cases.

1.3.4 Summary

Experiments reported in this section clearly demonstrated that combining morpho-syntactic knowledge in an ASR system at the sentence level is an efficient strategy, resulting in improved transcriptions and confidence measures. It is worthwhile to note that the combined score defined in (1.6) implements a linear combination of log-scores similar to score combination in multistream HMMs as discussed in chapter **??**. Moreover, we observed that the output after morpho-syntactic rescoring is more grammatical, a fact from which further NLP algorithms applied to the text resulting from the transcription should benefit.

1.4 Pragmatic knowledge integration

In this section, we present another step toward a better integration between ASR and NLP, focusing on pragmatic knowledge. In this framework, we consider topic-related information in order to adapt the LM of the ASR system in an unsupervised way.

Usually, LMs are trained once and for all on large multi-topic corpora. Every (part of the) document is then processed using the same general-purpose LM, whatever the actual topic, even though word frequencies depend on the theme. Topic-specific LMs are therefore a good way of improving ASR based on pragmatic knowledge. NLP methods precisely able to locate and characterize topics can be applied to update the vocabulary of an ASR system or its general-purpose LM [1, 4]. In this section, we focus on the adaptation of the LM, leaving the vocabulary untouched.

As presented in Fig 1.1, the basic idea of our approach is, first, to segment a broadcast transcript obtained with a baseline, general-purpose, LM into thematically coherent successive parts. For each segment, topic-specific data are then retrieved from a large collection of texts, i.e., the Internet, and used to modify the initial LM. To achieve this goal, an adaptation LM is obtained from the topic-specific data and linearly combined with the general-purpose LM thus resulting in an adapted LM. The latter is used to get a new, hopefully better, transcription for the corresponding segment. This adaptation process is repeated for each part of the document resulting from the segmentation step.

Note that for multimedia documents for which text data are already available, gathering topic-specific data can be done according to the available textual modality rather than based on a first transcription result. In the typology

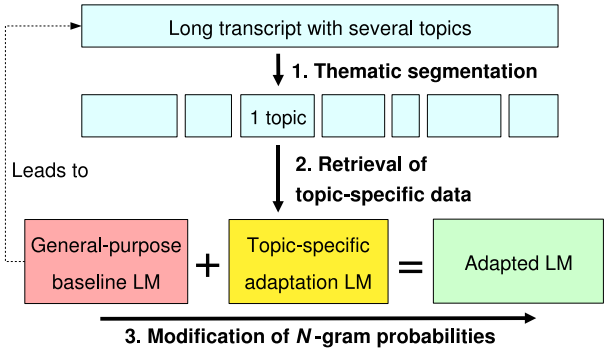

Fig. 1.1. Main scheme of a topic-based adaptation.

of fusion methods of Snoek and Worring [23], the proposed approach can be seen as an iterative fusion scheme where the audio and text modalities are considered sequentially. To achieve a better cascading of the modality, we present some adaptation of NLP algorithms to deal with the specificities of automatic transcriptions.

We first briefly review related works before presenting our approaches for the topic segmentation of transcriptions and for the creation of a topic-specific corpus from the Internet.

1.4.1 Related works

Most related works focus on only one subtask of the entire adaptation process—such as thematic segmentation, topic-specific data collecting or LM adaptation—and the combination of these subtasks as a whole topic-based LM adaptation process is still marginal [7].

The most popular indicator for the segmentation of texts into thematically consistent sections is lexical cohesion [29] which focuses on the vocabulary used in a text block and studies the numbers of word occurrences. Indeed, the frequent use of the same words in a given text section tends to demonstrate a thematic coherence of the text. This method can be enriched by the knowledge of more complex relations between words, such as synonymy. On top of lexical cohesion, other useful indicators of a topic change can be considered. For example, discourse markers [18], like "*however*" and "*furthermore*", can be used. In the case of multimedia documents, cues from the other modalities [28], e.g., shot boundaries, speaker changes or silences, provide valuable information.

Existing approaches for topic-specific data retrieval mainly differ according to the type of data collection used and the criterion chosen to select the relevant documents. Some studies are based on static sets of articles from which topic-specific texts can only be found for a restricted number of domains [14] while other, more recent, works seek to retrieve texts from the Internet [26].

This last source is more interesting, being an open resource which contains texts whose style is closer to speech than in typical written documents [31]. As for the method used to select topic-related texts, several criteria based on the word distributions can be considered to compare documents, as is classically done in the information retrieval (IR) domain [7].

Finally, language model adaptation given a topic-specific corpus of texts has been widely studied. A simple approach consists in training a LM from the corpus before interpolating the adaptation LM with the general-purpose one, either linearly or log-linearly. The N-gram probabilities of each LM are thus directly mixed. Other more complex techniques do not rely on an intermediate adaptation LM but rather search for a final N-gram distribution which minimizes an information quantity, like entropy or mutual information, according to constraints derived from the adaptation corpus. It has been shown that these methods outperform the interpolation-based ones [7].

As opposed to previous works, we study the complete adaptation process and propose a fully unsupervised approach, for which no restricting hypothesis on the domain or the number of topics is made. To this end, we first combine acoustic and text features for the segmentation of transcribed text. We then adapt NLP techniques to take into account confidence measures in order to gather topic-specific corpora from the Internet. Finally, we demonstrate that the cascaded use of NLP on transcriptions can benefit to the ASR system by providing adaptation data. The following sections describe each of these steps.

1.4.2 Transcript segmentation

Transcript segmentation is primarily based on the statistical lexical cohesion method described in [29]. In this method, a graph of all the possible segmentations is constructed where the vertex values represent the lexical cohesion for the segment represented by the vertex. Although originally designed for written documents, we observed that this method is quite robust to misrecognized words and segmentation into breath-groups [11]. However, the voluntary absence of word repetitions—for obvious stylistic reasons—limits the performance. We therefore extend semantic links between words by studying co-occurrences of lexical units in the French corpus *Le Monde*. On top of lexical cohesion, syntactic and acoustic cues were also considered. Syntactic cues are based on the sequences of words and POS tags to determine hidden boundaries between words. Moreover, as spoken documents are multimodal by nature, we take advantage of audio cues such as male/female speaker changes or jingles[7]. To accommodate the additional features, we extended the statistical lexical cohesion method by adapting the vertex weights to take into

[7] Surprisingly, pause duration turned out to be quite uninformative for the segmentation contrary to many previous studies on spoken document segmentation. This is mostly due to the nature of the documents, as radio broadcast news exhibit very few pauses.

account the syntactic and acoustic cues. To predict segment boundaries at the end of an hypothesized segment, a decision tree is used for the acoustic cues while a hidden N-gram models the syntactic information [11]. Vertex weights are modified so as to be a linear combination of the lexical cohesion, syntactic and acoustic log-scores.

Using only lexical cohesion leads to a recall of 57.4 % for a precision of 36.1 % on segment boundaries, resulting in 78.8 % of the segments containing a single topic. With the addition of semantic information plus syntactic and acoustic cues, we achieved a recall of 67.2 % with a precision of 43.2 %, yielding 83.5 % of pure segments.

To validate the other steps of our adaptation method without the bias of non thematically homogeneous segments, we consider manual topic segmentation in the rest of this section.

1.4.3 Language model adaptation

In order to train an adaptation LM for a thematically consistent section, keywords are automatically selected and submitted to a Web search engine, the resulting pages forming the adaptation corpus from which a LM is estimated. This adaptation LM is combined with the general-purpose baseline LM using linear interpolation, to obtain an adapted LM which is then used to rescore word graphs and generate a new transcription.

However, gathering an adaptation corpus from the transcription of a thematically homogeneous segment is far from trivial. First, keywords must be significant enough to fully characterize the content of the segment. On the other hand, too specific keywords are problematic as they usually result in few matches on the Internet. This remark raises questions about the "optimal" size of the adaptation corpus and its homogeneity. These many issues are discussed below.

Keyword spotting

Keywords are selected based on the well-known IR score *tf*idf*, where *tf* represents the frequency of a term w and *idf* is a value related to the inverse number of documents containing w in a text collection. Terms with the higher *tf*idf* scores are considered as characteristic terms and selected. In practice, the scores are computed on stems rather than words.

The standard *tf*idf* keyword selection method was adapted to take into account specificities of the documents at hand. First, proper names tend to result in very small and too specific adaptation corpora. A penalty is therefore applied to their *tf*idf* score which is scaled by a coefficient empirically set to 0.75. Because of the lack of cases in the transcribed texts, proper names are detected based on morpho-syntactic tags (see Section 1.3) combined with a dictionary: nouns with no definition in the dictionary are considered as proper names. Second, the *tf*idf* score of a term w is biased based on the confidence

measure c of w in order to limit the impact of misrecognized words, according to

$$score(w) = tf{*}idf \times \lambda + tf{*}idf \times (1 - \lambda) \times c \ , \tag{1.10}$$

where λ limits the influence of c.

Adaptation corpus creation

Even if the number of selected keywords is limited to five, combining them in a single query is not relevant for the task of gathering topic-related documents. Two main problems occur: a single query often results in very small amounts of adaptation data; moreover, the impact of transcription errors is detrimental. We rather rely on a fixed number of queries combining subsets of the whole keyword set. For example, a first query can be composed of the two best-scored keywords while the second one combines the first and third keywords. This strategy maximizes the probability of having at least one relevant query, even when transcription errors are present.

As queries can return several thousands of documents, the number of matching documents must be limited. In our study, it was experimentally observed that at least fifty documents are required to get a good adaptation LM. However, increasing the number of considered documents linearly increases runtime for a limited gain. Consequently, two hundred Web pages are kept. A cosine similarity distance between the initial transcription of the segment and each Web page is used to filter out irrelevant matches.

Results

Experiments were carried out on a subset of 22 manually selected segments from our broadcast news corpus. Perplexity before and after interpolation of the baseline LM with the one obtained from the adaptation data are reported in figure 1.2 for each of these segments. Perplexity measures how well a LM can predict the next word given the word history, where the lower the perplexity, the better. These results indicate that adaptation always reduces perplexity, even for texts with a low initial perplexity (texts 13, 19 and 20).

However, due to the complex interactions between all the components of an ASR system, decreasing the perplexity of a LM does not necessarily result a decrease of the word and sentence error rates. In two out of three segments, the perplexity falls by over 10 % which translates into a global absolute WER decrease of 0.2 %. This small global WER reduction is mostly due to the fact that the WER increases in 33 % of the sections while limited WER reductions are observed in the remaining ones.

Though mitigated, these first results are encouraging as they demonstrate the validity of the proposed unsupervised adaptation scheme. A detailed analysis of the transcriptions after adaptation shows that while topic-specific terms

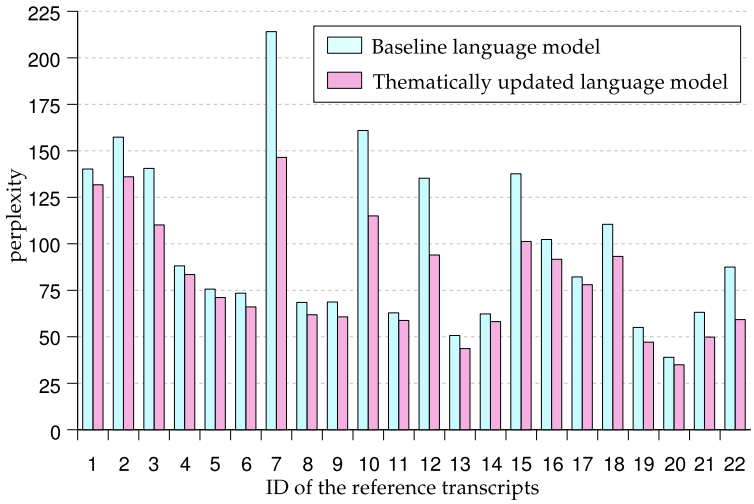

Fig. 1.2. Details of measured perplexity before and after our topic-based adaptation.

are better recognized, more new errors appear on grammatical words (prepositions, determiners, *etc.*). This can be partially explained by the fact that the adaptation LM is poor on grammatical words due to the limited size of the training corpus. We believe that better LM adaptation techniques than interpolation should be considered to circumvent this problem.

Summary

Even if segmentation has not been yet coupled with adaptation, this section illustrates the use of pragmatic information in combination with ASR system for unsupervised topic adaptation. The proposed approach mixes information from the text and audio modalities at various levels. For example, segmentation of transcriptions rely on lexical, syntactic, semantic and acoustic cues. Acoustic based confidence measures are used in the keyword selection process which is by itself based on the text modality. The sequential use of transcription, text analysis and again transcription is another example of multimodal fusion.

1.5 Conclusion

In this chapter, we presented experiments toward a better integration between automatic speech recognition and natural language processing techniques in order to improve multimedia (or spoken) document processing techniques based on a fusion of information from the audio and text modalities. In particular, we investigated the fusion of morpho-syntactic and of pragmatic

knowledge in an ASR system and demonstrated the benefits of it. We have seen that traditional multimodal fusion schemes such as the combination of log-scores or sequential processing of the modalities successfully apply to the text and audio modalities.

Many other research directions have to be investigated towards a full integration of these two modalities. For example, we have used $\mathcal{N}$-best lists at the interface between speech and natural language. This is convenient because each entry can be considered as a regular sentence thus enabling the use of standard NLP algorithms and a combination of knowledge sources at the sentence level. However, alternate transcription hypotheses are lost and NLP techniques can hardly recover from errors made by the ASR system. Using other interfaces, such as word graphs or confusion networks, might prove interesting but requires a deeper modification of standard NLP techniques. Finally, many other sources of linguistic knowledge not considered in this chapter can also benefit to ASR transcriptions, such as syntactic analysis or a more extensive use of semantic relations.

References

1. A. Allauzen and J.-L. Gauvain, "Open vocabulary ASR for audiovisual document indexation," in *Proc. IEEE Int'l Conf. Acous., Speech, and Signal Processing*, 2005.
2. F. Béchet, A. L. Gorin, J. H. Wright, and D. Hakkani-Tür, "Detecting and extracting named entities from spontaneous speech in a mixed initiative spoken dialogue context: How may I help you?" *Speech Communication*, vol. 42, no. 2, pp. 207–225, 2004.
3. J. R. Bellegarda, "Large vocabulary speech recognition with multispan statistical language models," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 1, pp. 76–84, 2000.
4. B. Bigi, Y. Huang, and R. De Mori, "Vocabulary and language model adaptation using information retrieval," in *Proc. Int'l Conf. on Spoken Language Processing*, 2004.
5. P. F. Brown, V. J. D. Pietra, P. V. de Souza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.
6. C. Chelba and F. Jelinek, "Structured language modeling," *Computer Speech and Language*, vol. 14, no. 4, pp. 283–332, 2000.
7. L. Chen, J.-L. Gauvain, L. Lamel, and G. Adda, "Dynamic language modeling for broadcast news," in *Proc. Int'l Conf. on Spoken Language Processing*, 2004.
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. Int'l Conf. on Speech Communication and Technology*, 2005.
9. D. Gildea and T. Hofmann, "Topic-based language models using EM," in *Proc. European Conf. on Speech Communication and Technology*, 1999.
10. P. A. Heeman, "POS tags and decision trees for language modeling," in *Proc. the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.
11. S. Huet, "Informations morpho-syntaxiques et adaptation thématique pour améliorer la reconnaissance de la parole," Ph.D. dissertation, Université de Rennes 1, Rennes, France, dec 2007.

12. S. Huet, G. Gravier, and P. Sébillot, "Morphosyntactic processing of N-best lists for improved recognition and confidence measure computation," in *Proc. Int'l Conf. on Speech Communication and Technology*, 2007.
13. R. Iyer and M. Ostendorf, "Modeling long distance dependence in language: Topic mixtures *versus* dynamic cache models," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 30–39, 1999.
14. D. Klakow, "Selecting articles from the language model training corpus," in *Proc. IEEE Int'l Conf. Acous., Speech, and Signal Processing*, 2000.
15. Y. Liu, E. Shriberg, A. Stolcke, and M. P. Harper, "Using machine learning to cope with imbalanced classes in natural speech: Evidence from sentence boundary and disfluency detection," in *Proc. Int'l Conf. on Spoken Language Processing*, 2004.
16. G. Maltese and F. Mancini, "An automatic technique to include grammatical and morphological information in a trigram-based statistical language model," in *Proc. IEEE Int'l Conf. Acous., Speech, and Signal Processing*, 1992.
17. 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.
18. D. Marcu, "The rhetorical parsing of unrestricted texts: A surface-based approach," *Computational Linguistics*, vol. 26, no. 3, pp. 395–448, 2000.
19. B. Merialdo, "Tagging English text with a probabilistic model," *Computational Linguistics*, vol. 20, no. 2, pp. 155–171, 1994.
20. A. Nasr, Y. Estève, F. Béchet, T. Spriet, and R. de Mori, "A language model combining N-grams and stochastic finite state automata," in *Proc. European Conf. on Speech Communication and Technology*, 1999.
21. B. Rueber, "Obtaining confidence measures from sentence probabilities," in *Proc. European Conf. on Speech Communication and Technology*, 1997.
22. S. Seneff, M. McCandless, and V. Zue, "Integrating natural language into the word graph search for simultaneous speech recognition and understanding," in *Proc. European Conf. on Speech Communication and Technology*, 1995.
23. C. G. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, January 2005.
24. A. Stolcke, Y. König, and M. Weintraub, "Explicit word error minimization in N-best list rescoring," in *Proc. European Conf. on Speech Communication and Technology*, 1997.
25. B. Suhm and A. Waibel, "Towards better language models for spontaneous speech," in *Proc. Int'l Conf. on Spoken Language Processing*, 1994.
26. M. Suzuki, Y. Kajiura, A. Ito, and S. Makino, "Unsupervised language model adaptation based on automatic text collection from WWW," in *Proc. Int'l Conf. on Speech Communication and Technology*, 2006.
27. C. Tillmann and H. Ney, "Word triggers and the EM algorithm," in *Proc. of the Workshop Computational Natural Language Learning (CoNLL)*, 1997, pp. 117–124.
28. G. Tür, D. Hakkani-Tür, A. Stolcke, and E. Shriberg, "Integrating prosodic and lexical cues for automatic topic segmentation," *Computational Linguistics*, vol. 21, no. 1, pp. 31–57, 2001.
29. M. Utiyama and H. Isahara, "A statistical model for domain-independent text segmentation," in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2001, pp. 499–506.

30. A. Valli and J. Véronis, "Étiquetage grammatical de corpus oraux : problèmes et perspectives," *Revue française de linguistique appliquée*, vol. 4, no. 2, pp. 113–133, 1999.
31. D. Vaufreydaz, M. Akbar, and J. Rouillard, "Internet documents: A rich source for spoken language modeling," in *Proc. IEEE Workshop Automatic Speech Recognition and Understanding*, 1999.
32. D. Vergyri, K. Kirchhoff, K. Duh, and A. Stolcke, "Morphology-based language modeling for arabic speech recognition," in *Proc. Int'l Conf. on Spoken Language Processing*, 2004.
33. M. Weintraub, Y. Aksu, S. Dharanipragada, S. Khudanpur, H. Ney, J. Prange, A. Stolcke, F. Jelinek, and E. Shriberg, "LM95 project report: Fast training and portability," Center for Language and Speech Processing, Johns Hopkins University, Tech. Rep., 1996.

Index