Improving Broadcast News Transcription with a Precision Grammar and Discriminative Reranking

Tobias Kaufmann, Thomas Ewender, Beat Pfister

Speech Processing Group, ETH Zurich, Switzerland

kaufmann@tik.ee.ethz.ch, ewendert@tik.ee.ethz.ch, pfister@tik.ee.ethz.ch

Abstract

We propose a new approach of integrating a precision grammar into speech recognition. The approach is based on a novel robust parsing technique and discriminative reranking. By reranking 100-best output of the LIMSI German broadcast news transcription system we achieved a significant reduction of the word error rate by 9.6% relative. To our knowledge, this is the first significant improvement for a real-world broad-domain speech recognition task due to a precision grammar.

Index Terms: speech recognition, precision grammar

1. Introduction

It has been recognized for a long time that speech recognition ought to benefit from sophisticated linguistic information [1]. In fact, language models based on statistical parsers have been shown to significantly improve speech recognition [2, 3, 4]. However, precision grammars such as head-driven phrase structure grammars [5] have hardly been utilized for this purpose. The primary role of precision grammars in the context of speech recognition is usually that of providing a basis for speech understanding [6, 7], and the respective systems are typically restricted to rather narrow domains.

This may be surprising considering that precision grammars are designed to accurately distinguish between correct and incorrect sentences. Unfortunately, precision grammars come with certain drawbacks. Parsing with precision grammars is computationally expensive and requires very detailed lexical information. A more fundamental issue is that of robustness. The utterances to be processed may not be grammatically correct, and even if they were, we would still face the problem that the coverage of precision grammars is rather low due to the large number of rare grammatical constructions.

In this study, we are interested in how and how much speech recognition can benefit from a precision grammar, leaving aside the issues of efficiency and scalability. We propose a robust approach of integrating a precision grammar into a speech recognition framework and report a significant reduction of the word error rate for a broad-domain task, namely broadcast news transcription. In contrast to our earlier work [8] we used a discriminative rather than a generative approach and we did not simplify the recognition task in any way. We also discuss some of the properties of our approach.

2. Approach

2.1. Architecture Overview

The overall architecture of our approach is shown in Figure 1. First, a speech signal is processed by a baseline speech recognizer that outputs a word lattice. As the speech signal in general covers more than one sentence, a segmentation component is used to guess the sentence boundaries. The word lattice is now segmented according to these sentence boundaries. This results in a number of sub-lattices, each corresponding to a sentence-like unit.

After this preprocessing step, each sub-lattice is processed as follows. First, the N best hypotheses are extracted together with their respective recognition scores or likelihoods. Next, each hypothesis is parsed exhaustively. The potential ambiguities are not resolved at this stage. Instead, all possible derivations are stored in a packed parse forest representation, which is disambiguated in the subsequent step. The result of disambiguation is a single parse tree for each hypothesis. Finally, the N hypotheses are compared to each other on the basis of their recognition scores as well as their respective parse trees. After this discriminative reranking procedure, the most likely hypothesis is chosen as the actual recognition result.

We have chosen this particular architecture as we did not want to artificially restrict the precision and complexity of the linguistic subsystem. Processing single hypotheses in isolation does not impose any additional constraints on the parser or the grammar, whereas a tighter coupling with the maximum a posterior (MAP) decoder would have implications with respect to both the grammar formalism and the resources available for parsing. In addition, this architecture provides a natural baseline, namely the word error rate of the first-best hypotheses, to which the extended system can be compared.

As for the segmentation, it has been shown that accurate sentence boundaries improve the performance of syntax-based language models [9]. Shorter segments are also preferrable from a non-linguistic point of view: as only the N best hypotheses are considered, splitting a long segment into M shorter ones potentially leads to N^M hypotheses to choose from.

Discriminative reranking of speech recognition hypotheses has been used before [4, 9]. However, these approaches were based on statistical parsers that produce a complete parse tree for any word sequence. As precision grammars are designed to reject incorrect word sequences, we had to deal with robustness in a completely different way.

2.2. Discriminative Reranking

Parse disambiguation and choosing the most likely hypothesis are both based on discriminative reranking with log-linear conditional models [10, 11, 12]. Given a set $\mathcal{Y}$ of candidates, the probability of a candidate $y \in \mathcal{Y}$ being the optimal choice is modeled as

$$P_\theta(y|\mathcal{Y}) = \frac{e^{\sum_j \theta_j f_j(y)}}{\sum_{y' \in \mathcal{Y}} e^{\sum_j \theta_j f_j(y')}} \qquad (1)$$
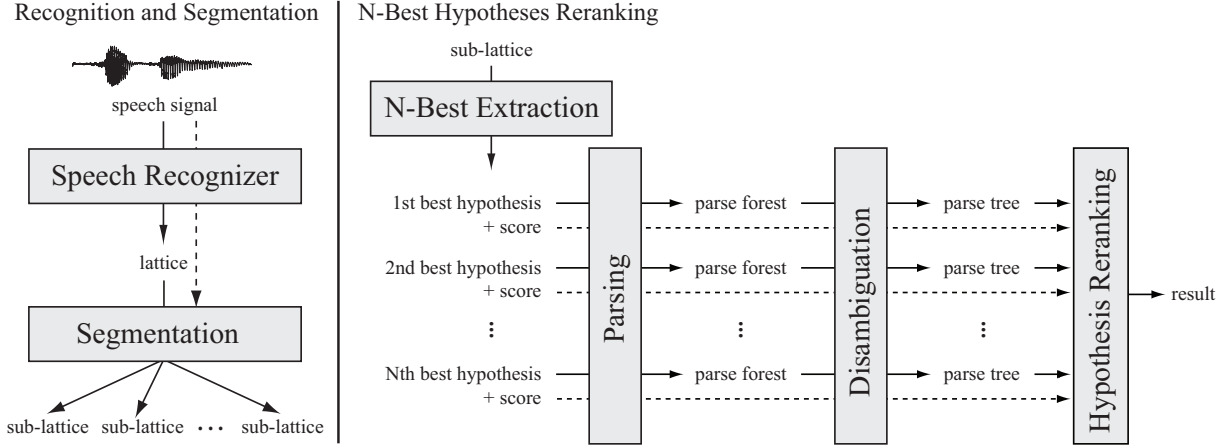
6 – 10 September, Brighton UK

Figure 1: *Architecture*

In the above equation, $\theta = (\theta_1, ..., \theta_n)$ denotes the vector of model parameters and $f_j(y), 1 \leq j \leq n$ are real-valued features. To estimate θ, we minimize the loss function $L(\theta)$ defined in [11]:

$$L(\theta) = -log \prod_s P_\theta(\mathcal{Y}^{(s)}_+ | \mathcal{Y}^{(s)}) + c \sum_j \theta_j^2 \qquad (2)$$

$$P_\theta(\mathcal{Y}_+ | \mathcal{Y}) = \sum_{y \in \mathcal{Y}_+} P_\theta(y | \mathcal{Y}) \qquad (3)$$

The set $\mathcal{Y}^{(s)}$ contains the candidates for a given training sample s, whereas $\mathcal{Y}^{(s)}_+ \subseteq \mathcal{Y}^{(s)}$ denotes the set of optimal candidates for sample s. The parameter c controls the amount of regularization and is used to avoid overfitting.

2.3. Robust Parsing

Precision grammars typically use discriminative reranking for parse disambiguation. In this context, the candidate set $\mathcal{Y}$ contains all parse trees y that can be derived from the given word sequence. A feature $f_j(y)$ counts how often a certain linguistic event (e.g. the presence of a certain grammatical construction) occurs in the parse tree y. Disambiguation now consists of choosing the parse tree which maximizes $P_\theta(y | \mathcal{Y})$. This is equivalent to maximizing the expression $\sum_j \theta_j f_j(y)$, which we subsequently refer to as the disambiguation score $s_d(y)$.

In real applications, it frequently occurs that the parser fails to derive a parse tree because the word sequence does not obey the grammar rules. Similarly, it can happen that the parser cannot derive the correct parse tree for some word sequence and therefore assigns it a highly implausible parse tree instead. In both cases, we would like the parser to identify the most plausible partial parse trees rather than producing no parse tree at all or a complete but implausible parse tree. We call this a robust parsing scheme.

A common approach to robust parsing is to create an artificial parse tree by choosing a sequence of partial parse trees (which can trivially be single words) and attaching them to a common root node. This shifts the problem to determining the optimal sequence of partial parse trees. Typical heuristics are to minimize the number of partial parse trees or to maximize the length of the largest partial parse tree [11, 13]. In the following, we will suggest a different approach that is tightly integrated into parse disambiguation.

Let the artificial parse tree $y^\star$ be composed of the partial parse trees $p_1, ..., p_k$. If we assume that the features $f_j(y)$ count linguistic events that are completely contained within the individual partial parse trees, $s_d(y^\star)$ is equal to $\sum_i s_d(p_i)$. Thus, the artificial parse tree $y^\star$ is optimal if and only if $\sum_i s_d(p_i)$ is maximal. The optimal artificial parse tree can easily be determined by means of dynamic programming.

To take the number of partial parse trees into account, we define a feature $f_p(y)$ which simply counts the partial parse trees in y. Note that the above assumption holds, as the event of a partial parse tree occurring is trivially contained within the partial parse trees. Similarly, we can define more specific features which count the partial parse trees with a certain type, e.g. adverbial or nominative noun phrase. Such features can help to model elliptical speech.

To some extent, it is also possible to handle events that are not contained within the partial parse trees. For example, we could define a feature $f_m(y)$ which is 1 if and only if y contains two or more partial parse trees. To deal with such features, additional state information (in this case the number of partial parse trees) has to be included in the dynamic programming.

To train the disambiguation model, we construct candidate sets $\mathcal{Y}^{(s)}$ containing the optimal (artificial) parse tree together with a number of randomly sampled parse trees with different numbers of partial parse trees.

2.4. Features

For disambiguation we used a set of more than 12,000 features that count specific local configurations in the parse tree. The vast majority of these features are instances of generic feature templates and count events such as sequences of consecutive rule applications. Other features were specifically designed to count constructions such as certain types of noun phrases. There is also a set of features that are based on the output of a part-of-speech tagger. These features count mismatches between the predicted part-of-speech tags and the ones actually occurring in the parse tree. Finally, we used features that depend on the number and type of partial parse trees as introduced in the previous section.

For hypothesis reranking, we used all of the above disambiguation features together with a set of additional features. In particular, there are features representing the acoustic score, the weighted language model score as well as the total recognition

score which is just the sum of the former two. Another feature represents the disambiguation score of the parse tree. Finally, there is a set of features counting the number of partial parse trees that border a prosodic boundary of a certain strength. The strength of a prosodic boundary is quantified as the posterior probability of a sentence boundary at the given position. This information is provided by the segmentation component.

2.5. Miscellaneous Components

We have developed a segmentation system along the lines of [14]. For every inter-word boundary of the first-best hypothesis, the segmentation system estimates the posterior probability of a sentence boundary occurring. A segment boundary is set at every position where this probability is greater than 0.5. Segments with more than 25 words are split at the position with the highest sentence boundary probability.

To compute the part-of-speech features, we have developed a maximum entropy part-of-speech tagger tailored to speech recognizer output. The tagger was trained on the the German TIGER corpus [15], which was previously transformed into a format that resembles the output of the baseline speech recognizer.

The parser was developed for precursory experiments and is described in [8] and [16]. The discriminative reranking models were trained with the reranker software presented in [12].

3. Experiments

3.1. Data

Our experiments are based on word lattice output of the LIMSI German broadcast news transcription system [17]. Three of the six available news shows were already used in our earlier experiments. We therefore did not consider these data except for the training of our segmentation system. The remaining three news shows were automatically segmented into 603 sentence-like units. For each unit, the 100 best recognition hypotheses were extracted. All words occurring in the 100-best lists were collected in a single set. This subset of the speech recognizer vocabulary was the basis on which the linguistic resources (see Section 3.2) were developed and refined.

For the training of the disambiguation model, the reference transcription of each segment was manually annotated with its syntactic structure. These data were complemented with a tree-bank of about 440 sentences that was created for earlier experiments.

Inspection of the data reveals that for 94% of the segments the correct hypothesis is among the 100 best hypotheses. The word error rate of the first-best hypotheses is 13.27%. By always choosing the optimal (but not necessarily correct) hypothesis, an oracle word error rate of 6.32% can be achieved.

3.2. Linguistic Resources

The grammar and the lexicon were both developed without any knowledge of the test data except for the set of the roughly 7000 words occurring in the N-best lists. The difficulties in acquiring the lexicon were twofold. First, every lexicon entry had to be annotated with detailed syntactic information which could not be directly derived from existing resources (e.g. subcategorization frames for verbs). Second, lexemes consisting of two or more words (e.g. *"by and large"* in English) had to be identified. We used statistical corpus-based techniques to assist the lexicon developer in both tasks.

Table 1: *Results*

approach	word error rate
baseline system	13.27%
automatic segmentation	11.98% (**-9.6% relative**)
manual segmentation	11.67% (-12.2% relative)
100-best oracle	6.32%

The German grammar used in this work is an instance of the head-driven phrase structure grammar formalism. The greater part of the grammar was developed for precursory experiments and is described in [8]. For the present experiment, we systematically extended the grammar by parsing newspaper text and analyzing the parse failures. The coverage of the current grammar is outlined by the collection of test sentences available at http://www.tik.ee.ethz.ch/~spr/hpsg0409/. In order to assess how well the grammar covers the given data, we parsed the manually segmented reference transcriptions. The correct parse tree could be derived for 61% of the sentences.

3.3. Setup

In order to make the best use of the available data, we followed a 10-fold cross-validation scheme: the hypothesis reranking model was tested on each fold after being trained on the remaining 9 folds.

This procedure was complicated by the fact that training and testing the hypothesis reranking model requires the most likely parse tree of each hypothesis in the training set and the test set. The extraction of the most likely parse tree in turn requires a well-trained disambiguation model. Thus, the disambiguation model was trained and applied by means of an embedded cross-validation cycle. To disambiguate the test set, the disambiguation model was simply trained on the training set. To disambiguate a fold of the training set, the disambiguation model was trained on the remaining 8 folds of the training set. This ensured that no information from the test set was used for training.

The regularization parameter c was set to 13 for disambiguation and 30 for hypothesis reranking. These values were found to work well in earlier experiments with similar data and feature sets.

3.4. Results

The results are shown in Table 1. By applying automatic segmentation and 100-best reranking we reduced the word error rate by 9.6% relative. This improvement is statistically significant on a level of less than 0.1% with respect to both the McNemar test and the Matched Pairs Sentence-Segment Word Error test [18]. If the same experiment is performed with manual segmentation, the word error rate is reduced by 12.2% relative. This is in line with the findings of [9] who observe that syntax-based language models benefit from accurate sentence boundaries.

For the above experiments, the number of hypotheses N was fixed to a value of 100. In order to assess the impact of this parameter on the word error rate, we repeated the experiment with automatic segmentation for all N in 1..100. We did not try higher values of N as the lexicon is only guaranteed to cover the words of the 100 best hypotheses. The results are shown in Figure 2. The 10 best hypotheses alone account for an improvement of about 1% absolute (7.8% relative). For higher values of
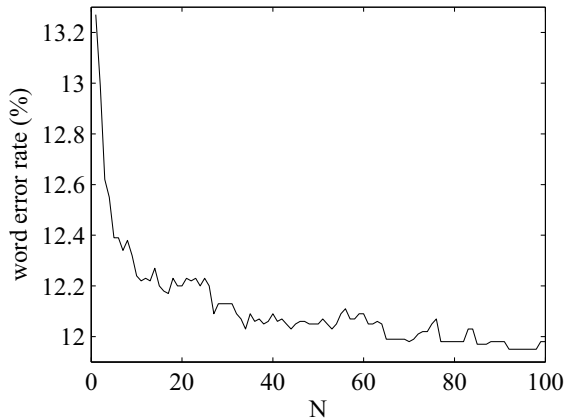
Figure 2: *Influence of the number of hypotheses N on the word error rate.*

N the word error rate starts to level off, but there still seems to be room for further improvement beyond N=100.

We were also interested in how well our approach performs for a baseline system with a lower word error rate. We modified the experiment with automatic segmentation by removing those 97 segments that are part of interviews or sports reports. As a result, the baseline word error rate dropped to 10.91%. The extended system achieved an improvement of 1.5% absolute or 13.7% relative. This confirms the intuition that a language model which makes extensive use of non-local information is particularly good at correcting errors within contexts that are largely correct.

We carried out further investigations on the original experiment with automatic segmentation. We observed that about 37% of the net error corrections were achieved by choosing a hypothesis with more than one partial parse tree. This indicates that our approach does exhibit certain robustness properties. As German is a highly inflected language, we further examined to what degree our syntax-based approach "simply" corrects inflectional endings rather than introducing an uninflected word or a stem that was not present in the first-best hypothesis. We found that less than 30% of the net error corrections were due to changing inflectional endings.

4. Conclusions

We have shown that the constraints encoded in a precision grammar can significantly improve speech recognition for a broad-domain task. We have also given evidence that our approach exhibits certain robustness properties.

The reported results were obtained with relatively little training data. The hypothesis reranking model was trained on speech recognizer output for about 550 sentence segments, and the treebank for training the disambiguation model contained roughly 1000 sentences or 12000 words. We suppose that this small amount of training data was sufficient because of the rich linguistic information encoded in the grammar and the lexicon.

5. Acknowledgements

6. References

[1] E. Brill, R. Florian, J. Henderson, and L. Mangu, "Beyond n-grams: can linguistic sophistication improve language modeling?" in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, 1998, pp. 186–190.

[2] C. I. Chelba, "Exploiting syntactic structure for natural language modeling," Ph.D. dissertation, 2000.

[3] K. Hall and M. Johnson, "Language modeling using efficient best-first bottom-up parsing," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2003, pp. 507–512.

[4] M. Collins, B. Roark, and M. Saraclar, "Discriminative syntactic language modeling for speech recognition," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, USA, 2005, pp. 507–514.

[5] C. J. Pollard and I. A. Sag, *Head-Driven Phrase Structure Grammar*. University of Chicago Press, 1994.

[6] G. V. Noord, G. Bouma, R. Koeling, and M.-J. Nederhof, "Robust grammatical analysis for spoken dialogue systems," *Natural Language Engineering*, vol. 5, no. 1, pp. 45–93, 1999.

[7] B. Kiefer, H.-U. Krieger, and M.-J. Nederhof, "Efficient and robust parsing of word hypotheses graphs," in *Verbmobil. Foundations of Speech-to-Speech Translation*, W. Wahlster, Ed. Springer, Berlin, 2000, pp. 280–295.

[8] T. Kaufmann and B. Pfister, "Applying a grammar-based language model to a broadcast-news transcription task," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, Columbus, USA*, June 2008.

[9] W. McNeill, J. Kahn, D. Hillard, and M. Ostendorf, "Parse structure and segmentation for improving speech recognition," in *IEEE Spoken Language Technology Workshop*, 2006, pp. 90–93.

[10] M. Johnson, S. Geman, S. Canon, Z. Chi, and S. Riezler, "Estimators for stochastic "unification-based" grammars," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, Morristown, NJ, USA, 1999, pp. 535–541.

[11] S. Riezler, T. King, R. Kaplan, R. Crouch, J. Maxwell III, and M. Johnson, "Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques," *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 271–278, 2002.

[12] E. Charniak and M. Johnson, "Coarse-to-fine n-best parsing and maxent discriminative reranking," *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 173–180, 2005.

[13] R. Prins and G. van Noord, "Reinforcing parser preferences through tagging," *Traitement Automatique des Langues*, vol. 44, no. 3, pp. 121–139, 2003.

[14] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[15] S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith, "The TIGER treebank," in *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria, 2002.

[16] T. Kaufmann and B. Pfister, "Applying licenser rules to a grammar with continuous constituents," in *Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar*, Stanford, USA, 2007, pp. 150–162.

[17] K. McTait and M. Adda-Decker, "The 300k LIMSI German broadcast news transcription system," in *Proceedings of Eurospeech*, Geneva, Switzerland, 2003.

[18] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceedings of the ICASSP*, 1989, pp. 532–535.