WordNet-based Semantic Relatedness Measures in Automatic Speech Recognition for Meetings

Michael Pucher
Telecommunications Research Center Vienna
Vienna, Austria
Speech and Signal Processing Lab, TU Graz
Graz, Austria
pucher@ftw.at

Abstract

This paper presents the application of WordNet-based semantic relatedness measures to *Automatic Speech Recognition* (ASR) in multi-party meetings. Different word-utterance context relatedness measures and utterance-coherence measures are defined and applied to the rescoring of N-best lists. No significant improvements in terms of *Word-Error-Rate* (WER) are achieved compared to a large word-based n-gram baseline model. We discuss our results and the relation to other work that achieved an improvement with such models for simpler tasks.

1 Introduction

As (Pucher, 2005) has shown different WordNet-based measures and contexts are best for word prediction in conversational speech. The JCN (Section 2.1) measure performs best for nouns using the noun-context. The LESK (Section 2.1) measure performs best for verbs and adjectives using a mixed word-context.

Text-based semantic relatedness measures can improve word prediction on simulated speech recognition hypotheses as (Demetriou et al., 2000) have shown. (Demetriou et al., 2000) generated N-best lists from phoneme confusion data acquired from a speech recognizer, and a pronunciation lexicon. Then sentence hypotheses of varying *Word-Error-Rate* (WER) were generated based on sentences from different genres from the *British National Corpus* (BNC). It was shown by them that the semantic model can improve recognition, where the amount of improvement varies with context length and sentence length. Thereby it was shown that these models can make use of long-term information.

In this paper the best performing measures from (Pucher, 2005), which outperform baseline models on word prediction for conversational telephone speech are used for *Automatic Speech Recognition* (ASR) in multi-party meetings. Thereby we want to investigate if WordNet-based models can be used for rescoring of ‘real’ N-best lists in a difficult task.

1.1 Word prediction by semantic similarity

The standard n-gram approach in language modeling for speech recognition cannot cope with long-term dependencies. Therefore (Bellegarda, 2000) proposed combining n-gram language models, which are effective for predicting local dependencies, with *Latent Semantic Analysis* (LSA) based models for covering long-term dependencies. WordNet-based semantic relatedness measures can be used for word prediction using long-term dependencies, as in this example from the CallHome English telephone speech corpus:

(1) B: I I well, you should see what the $\lfloor$students$\rfloor$
B: after they torture them for six $\lfloor$years$\rfloor$ in middle $\lfloor$school$\rfloor$ and high $\lfloor$school$\rfloor$ they don’t want to do anything in $\lfloor$college$\rfloor$ particular.

In Example 1 *college* can be predicted from the noun context using semantic relatedness measures,

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 129–132,
Prague, June 2007.

here between *students* and *college*. A 3-gram model gives a ranking of *college* in the context of *anything in*. An 8-gram predicts *college* from *they don't want to do anything in*, but the strongest predictor is *students*.

1.2 Test data

The JCN and LESK measure that are defined in the next section are used for N-best list rescoring. For the WER experiments N-best lists generated from the decoding of conference room meeting test data of the NIST Rich Transcription 2005 Spring (RT-05S) meeting evaluation (Fiscus et al., 2005) are used. The 4-gram that has to be improved by the WordNet-based models is trained on various corpora from conversational telephone speech to web data that together contain approximately 1 billion words.

2 WordNet-based semantic relatedness measures

2.1 Basic measures

Two similarity/distance measures from the Perl package WordNet-Similarity written by (Pedersen et al., 2004) are used. The measures are named after their respective authors. All measures are implemented as similarity measures. JCN (Jiang and Conrath, 1997) is based on the information content, and LESK (Banerjee and Pedersen, 2003) allows for comparison across Part-of-Speech (POS) boundaries.

2.2 Word context relatedness

First the relatedness between words is defined based on the relatedness between senses. $\mathrm{S}(w)$ are the senses of word w. Definition 2 also performs word-sense disambiguation.

$$\mathrm{rel}(w, w') = \max_{c_i \in \mathrm{S}(w)\ c_j \in \mathrm{S}(w')} \mathrm{rel}(c_i, c_j) \qquad (2)$$

The relatedness of a word and a context (rel_W) is defined as the average of the relatedness of the word and all words in the context.

$$\mathrm{rel}_\mathrm{W}(w, C) = \frac{1}{\mid C \mid} \sum_{w_i \in C} \mathrm{rel}(w, w_i) \qquad (3)$$

2.3 Word utterance (context) relatedness

The performance of the word-context relatedness (Definition 3) shows how well the measures work for algorithms that proceed in a left-to-right manner, since the context is restricted to words that have already been seen. For the rescoring of N-best lists it is not necessary to proceed in a left-to-right manner. The word-utterance-context relatedness can be used for the rescoring of N-best lists. This relatedness does not only use the context of the preceding words, but the whole utterance.

Suppose $U = \langle w_1, \ldots, w_n \rangle$ is an utterance. Let $\mathrm{pre}(w_i, U)$ be the set $\bigcup_{j<i} w_j$ and $\mathrm{post}(w_i, U)$ be the set $\bigcup_{j>i} w_j$. Then the word-utterance-context relatedness is defined as

$$\mathrm{rel}_{\mathrm{U}_1}(w_i, U, C) = \mathrm{rel}_\mathrm{W}(w_i, \mathrm{pre}(w_i, U) \cup \mathrm{post}(w_i, U) \cup C)\ . \qquad (4)$$

In this case there are two types of context. The first context comes from the respective meeting, and the second context comes from the actual utterance.

Another definition is obtained if the context C is eliminated ($C = \emptyset$) and just the utterance context U is taken into account.

$$\mathrm{rel}_{\mathrm{U}_2}(w_i, U) = \mathrm{rel}_\mathrm{W}(w_i, \mathrm{pre}(w_i, U) \cup \mathrm{post}(w_i, U)) \qquad (5)$$

Both definitions can be modified for usage with rescoring in a left-to-right manner by restricting the contexts only to the preceding words.

$$\mathrm{rel}_{\mathrm{U}_3}(w_i, U, C) = \mathrm{rel}_\mathrm{W}(w_i, \mathrm{pre}(w_i, U) \cup C) \qquad (6)$$

$$\mathrm{rel}_{\mathrm{U}_4}(w_i, U) = \mathrm{rel}_\mathrm{W}(w_i, \mathrm{pre}(w_i, U)) \qquad (7)$$

2.4 Defining utterance coherence

Using Definitions 4-7 different concepts of utterance coherence can be defined. For rescoring the utterance coherence is used, when a score for each element of an N-best list is needed. U is again an utterance $U = \langle w_1, \ldots, w_n \rangle$.

$$\mathrm{cohU}_1(U,C) = \frac{1}{\mid U \mid} \sum_{w \in U} \mathrm{rel}_{\mathrm{U}_1}(w,U,C) \quad (8)$$

The first semantic utterance coherence measure (Definition 8) is based on all words in the utterance as well as in the context. It takes the mean of the relatedness of all words. It is based on the word-utterance-context relatedness (Definition 4).

$$\mathrm{cohU}_2(U) = \frac{1}{\mid U \mid} \sum_{w \in U} \mathrm{rel}_{\mathrm{U}_2}(w,U) \quad (9)$$

The second coherence measure (Definition 9) is a pure inner-utterance-coherence, which means that no history apart from the utterance is needed. Such a measure is very useful for rescoring, since the history is often not known or because there are speech recognition errors in the history. It is based on Definition 5.

$$\mathrm{cohU}_3(U,C) = \frac{1}{\mid U \mid} \sum_{w \in U} \mathrm{rel}_{\mathrm{U}_3}(w,U,C) \quad (10)$$

The third (Definition 10) and fourth (Definition 11) definition are based on Definition 6 and 7, that do not take future words into account.

$$\mathrm{cohU}_4(U) = \frac{1}{\mid U \mid} \sum_{w \in U} \mathrm{rel}_{\mathrm{U}_4}(w,U) \quad (11)$$

3 Word-error-rate (WER) experiments

For the rescoring experiments the first-best element of the previous N-best list is added to the context. Before applying the WordNet-based measures, the N-best lists are POS tagged with a decision tree tagger (Schmid, 1994). The WordNet measures are then applied to verbs, nouns and adjectives. Then the similarity values are used as scores, which have to be combined with the language model scores of the N-best list elements.

The JCN measure is used for computing a noun score based on the noun context, and the LESK measure is used for computing a verb/adjective score based on the noun/verb/adjective context. In the end there is a *lesk_score* and a *jcn_score* for each N-best list. The final WordNet score is the sum of the two scores.

The log-linear interpolation method used for the rescoring is defined as

$$p(S) \propto p_{\mathrm{wordnet}}(S)^{\lambda}\, p_{n\text{-gram}}(S)^{1-\lambda} \quad (12)$$

where $\propto$ denotes normalization. Based on all WordNet scores of an N-best list a probability is estimated, which is then interpolated with the n-gram model probability. If only the elements in an N-best list are considered, log-linear interpolation can be used since it is not necessary to normalize over all sentences. Then there is only one parameter λ to optimize, which is done with a brute force approach. For this optimization a small part of the test data is taken and the WER is computed for different values of λ.

As a baseline the n-gram mixture model trained on all available training data (≈ 1 billion words) is used. It is log-linearly interpolated with the WordNet probabilities. Additionally to this sophisticated interpolation, solely the WordNet scores are used without the n-gram scores.

3.1 WER experiments for inner-utterance coherence

In this first group of experiments Definitions 8 and 9 are applied to the rescoring task. Similarity scores for each element in an N-best list are derived according to the definitions. The first-best element of the last list is always added to the context. The context size is constrained to the last 20 words. Definition 8 includes context apart from the utterance context, Definition 9 only uses the utterance context.

No improvement over the n-gram baseline is achieved for these two measures. Neither with the log-linearly interpolated models nor with the WordNet scores alone. The differences between the methods in terms of WER are not significant.

3.2 WER experiments for utterance coherence

In the second group of experiments Definitions 10 and 11 are applied to the rescoring task. There is again one measure that uses dialog context (10) and one that only uses utterance context (11).

Also for these experiments no improvement over the n-gram baseline is achieved. Neither with the log-linearly interpolated models nor with the WordNet scores alone. The differences between the methods in terms of WER are also not significant. There are also no significant differences in performance between the second group and the first group of experiments.

4 Summary and discussion

We showed how to define more and more complex relatedness measures on top of the basic relatedness measures between word senses.

The LESK and JCN measures were used for the rescoring of N-best lists. It was shown that speech recognition of multi-party meetings cannot be improved compared to a 4-gram baseline model, when using WordNet models.

One reason for the poor performance of the models could be that the task of rescoring simulated N-best lists, as presented in (Demetriou et al., 2000), is significantly easier than the rescoring of 'real' N-best lists. (Pucher, 2005) has shown that WordNet models can outperform simple random models on the task of word prediction, in spite of the noise that is introduced through word-sense disambiguation and POS tagging. To improve the word-sense disambiguation one could use the approach proposed by (Basili et al., 2004).

In the above WER experiments a 4-gram baseline model was used, which was trained on nearly 1 billion words. In (Demetriou et al., 2000) a simpler baseline has been used. 650 sentences were used there to generate sentence hypotheses with different WER using phoneme confusion data and a pronunciation lexicon. Experiments with simpler baseline models ignore that these simpler models are not used in today's recognition systems.

We think that these prediction models can still be useful for other tasks where only small amounts of training data are available. Another possibility of improvement is to use other interpolation techniques like the maximum entropy framework. WordNet-based models could also be improved by using a trigger-based approach. This could be done by not using the whole WordNet and its similarities, but defining word-trigger pairs that are used for rescoring.

5 Acknowledgements

This work was supported by the European Union 6th FP IST Integrated Project AMI (Augmented Multi-party Interaction, and by Kapsch Carrier-Com AG and Mobilkom Austria AG together with the Austrian competence centre programme **K***plus*.

References

Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th Int. Joint Conf. on Artificial Intelligence*, pages 805–810, Acapulco.

Roberto Basili, Marco Cammisa, and Fabio Massimo Zanzotto. 2004. A semantic similarity measure for unsupervised semantic tagging. In *Proc. of the Fourth International Conference on Language Resources and Evaluation (LREC2004)*, Lisbon, Portugal.

Jerome Bellegarda. 2000. Large vocabulary speech recognition with multispan statistical language models. *IEEE Transactions on Speech and Audio Processing*, 8(1), January.

G. Demetriou, E. Atwell, and C. Souter. 2000. Using lexical semantic knowledge from machine readable dictionaries for domain independent language modelling. In *Proc. of LREC 2000, 2nd International Conference on Language Resources and Evaluation*.

Jonathan G. Fiscus, Nicolas Radde, John S. Garofolo, Audrey Le, Jerome Ajot, and Christophe Laprun. 2005. The rich transcription 2005 spring meeting recognition evaluation. In *Rich Transcription 2005 Spring Meeting Recognition Evaluation Workshop*, Edinburgh, UK.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, Taiwan.

Ted Pedersen, S. Patwardhan, and J. Michelizzi. 2004. WordNet::Similarity - Measuring the relatedness of concepts. In *Proc. of Fifth Annual Meeting of the North American Chapter of the ACL (NAACL-04)*, Boston, MA.

Michael Pucher. 2005. Performance evaluation of WordNet-based semantic relatedness measures for word prediction in conversational speech. In *IWCS 6, Sixth International Workshop on Computational Semantics*, Tilburg, Netherlands.

H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, September.