Using Domain Knowledge about Medications to Correct Recognition Errors in Medical Report Creation

Stephanie Schreitter
Alexandra Klein
Johannes Matiasek
Austrian Research Institute
for Artificial Intelligence (OFAI)
Freyung 6/6
1010 Vienna, Austria

Harald Trost
Section for Artificial Intelligence
Center for Med. Statistics, Informatics,
and Intelligent Systems
Medical University of Vienna
Freyung 6/2
1010 Vienna, Austria

firstname.lastname@ofai.at harald.trost@meduniwien.ac.at

Abstract

We present an approach to analysing automatic speech recognition (ASR) hypotheses for dictated medical reports based on background knowledge. Our application area is prescriptions of medications, which are a frequent source of misrecognitions: In a sample report corpus, we found that about 40% of the active substances or trade names and dosages were recognized incorrectly. In about 25% of these errors, the correct string of words was contained in the word graph. We have built a knowledge base of medications based on information contained in the Unified Medical Language System (UMLS), consisting of trade names, active substances, strengths and dosages. From this, we generate a variety of linguistic realizations for prescriptions. Whenever an inconsistency in a prescription is encountered on the best path of the word graph, the system searches for alternative paths which contain valid linguistic realizations of prescriptions consistent with the knowledge base. If such a path exists, a new concept edge with a better score is added to the word graph, resulting in a higher plausibility for this reading. The concept edge can be used for rescoring the word graph to obtain a new best path. A preliminary evaluation led to encouraging results: in nearly half of the cases where the word graph contained the correct variant, the correction was successful.

1 Introduction

Automatic speech recognition (ASR) is widely used in the domain of medical reporting. Users appreciate the fact that the records can be accessed immediately after their creation and that speech recognition provides a hands-free input mode, which is important as physicians often simultaneously handle documents such as notes and X-rays (Alapetite et al., 2009). A drawback of using ASR is the fact that speech-recognition errors have to be corrected manually by medical experts before the resulting texts can be used for electronic patient records, quality control and billing purposes. This manual post-processing is time-consuming, which slows down hospital workflows.

A number of recognition errors could be avoided by incorporating explicit domain knowledge. We consider prescriptions of medications a good starting point as they are common and frequent in the various medical fields. Furthermore, they contain trade names and dosages, i.e. proper names and digits, which are frequently misrecognized by ASR in all domains.

For our approach, we have extracted and adapted information about medications from the Unified Medical Language System (UMLS) (Lindberg et al., 1993). This data contains information about trade names, active substances, strengths and dosages and can easily be modified, e.g. when new medications are released.

In the first step, we assessed the potential for improvement by analyzing a sample corpus of medical reports. It turned out that in 4383 dictated reports which were processed by a speech-recognition system, the word-error rate for medications was about 40%, which is slightly higher than the the average word-error rate of the reports. Examining a sample

Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, pages 22–28,
Los Angeles, California, June 2010.

of word graphs for the reports, we realized that in about 30% of these errors, the correct string of words was contained in the word graph, but not ranked as the best path.

In the following sections, we will first give an overview of previous approaches to detecting speech-recognition errors and semantic rescoring of word-graph hypotheses. Then, we will describe how we have adapted information about medications from the UMLS to enhance the word graph with concept nodes representing domain-specific information. Finally, we will illustrate the potential for improving the speech-recognition result by means of an evaluation of word graphs for medical reports which were processed by our system.

2 Extraction of Medication Information, Error Handling and Semantic Rescoring

(Gold et al., 2008) gives an overview on extracting structured medication information from clinical narratives. Extracted medication information may serve as a base for quality control, pharmaceutical research and the automatic creation of Electronic Health Records (EHR) from clinical narratives. The *i2b2* Shared Task 2009 focussed on medication extraction, e.g. (Patrick and Li, 2009; Halgrim et al., 2010). These approaches work on written narrative texts from clinical settings, which may have been typed by physicians, transcribed by medical transcriptionists or recognized by ASR and corrected by medical transcriptionists.

In contrast, our approach takes as input word graphs produced by an ASR system from dictated texts and aims at minimizing the post-processing required by human experts.

Speech-recognition systems turn acoustic input into word graphs, which are directed acyclic graphs representing the recognized spoken forms and their confidence scores (Oerder and Ney, 1993). In most speech-recognition systems, meaning is implicitly represented in the language model (LM), indicating the plausibility of sequences of words in terms of n-grams. It has often been stated that the introduction of an explicit representation of the utterance meaning will improve recognition results. Naturally, this works best in limited domains: the larger an application domain, the more difficult it is to build an optimal knowledge representation for all possible user utterances. Limited domains seem to be more rewarding with regard to coverage and performance. Consequently, combining speech recognition and speech understanding has so far mostly resulted in applications in the field of dialogue systems where knowledge about the domain is represented in terms of the underlying database, e.g. (Seneff and Polifroni, 2000).

Several approaches have investigated the potential of improving the mapping between the user utterance and the underlying database by constructing a representation of the utterance meaning. Meaning analysis is either a separate post-processing step or an integral part of the recognition process. In some approaches, the recognition result is analyzed with regards to content to support the dialogue manager in dealing with inconsistencies (Macherey et al., 2003). As far as dictated input is concerned, which is not controlled by a dialogue manager, (Voll, 2006) developed a post-ASR error-detection mechanism for radiology reports. The hybrid approach uses statistical as well as rule-based methods. The knowledge source UMLS is employed for measuring the semantic distance between concepts and for assessing the coherence of the recognition result.

In other approaches, the analysis of meaning is integrated into the recognition process. Semantic confidence measurement annotates recognition hypotheses with additional information about their assumed plausibility based on semantic scores (Zhang and Rudnicky, 2001; Sarikaya et al., 2003). (Gurevych and Porzel, 2003; Gurevych et al., 2003) present a rescoring approach where the hypotheses in the word graph are reordered according to semantic information. Usually, conceptual parsers are employed which construct a parse tree of concepts representing the input text for mapping between the recognition result and the underlying representation. Semantic language modeling (Wanget al., 2004; Buehler et al., 2005) enhances the language model to incorporate sequences of concepts which are considered coherent and typical for a specific context. In these approaches, the representations of the underlying knowledge are created specially for the applications or are derived from a text corpus.

In our approach, we aim at developing a prototype

for integrating available knowledge sources into the analysis of the word graph during the recognition process. We have decided not to integrate the component directly into the ASR system but to introduce a separate post-processing step for the recognition of information about medications with the word graphs as interface. This makes it easier to update the medication knowledge base, e.g. if new medications are released. Furthermore, it is not necessary to retrain the ASR system language model for each new version of the medication knowledge base.

3 Knowledge Base and Text Corpus

For our approach, we prepared a knowledge base concerning medications and dosages, and we used a corpus of medical reports, dictated by physicians in hospitals. The ASR result and a manual transcription is available for each report. For a subset of the corpus, word graphs could be obtained. By aligning the recognition result with the manual transcriptions, error regions can be extracted.

3.1 Knowledge Base

As it is our aim to find correct dosages of medications in the word graph, we built a domain-specific knowledge base which contains medications and strengths as they occur in prescriptions. In our sample of medical reports, about 1/3 of the medications occurred as active ingredients while the rest were trade names. Therefore, both had to be covered in our knowledge base which is based on RxNorm (Liu et al., 2005). RxNorm is a standardized nomenclature for clinical drugs and drug delivery devices and part of UMLS, ensuring a broad coverage of trade names and active ingredients. Of several available versions of RxNorm, the semantic branded drug form is the most suitable one for our purposes as it contains pharmaceutical ingredients, strengths, and trade names. For example, the trade name *Synthroid*® is listed as follows:

Thyroxine 0.025 MG Oral Tablet [Synthroid®]

Thyroxine is the active ingredient with the dosage value 0.025 and the dosage unit milligrams. The dosage unit form is oral tablet.

We used a RxNorm version with 1,508 active substances and 7,688 trade names (11,263 trade names counting the different dosages). The active ingredients in RxNorm are associated with Anatomical Therapeutic Chemical (ATC) Codes.

3.2 Sample Corpus

The corpus is a random sample of 924 clinical reports which were dictated by physicians from various specialties and hospitals. The dications were processed by an ASR system and transcribed by human experts. Word graphs marked with the best path (indicating the highest acoustic and language-model scores) represent the recognition result. Tradenames are part of the recognition lexicon, but they are frequently misrecognized.

Of the 9196 medications (i.e. trade names and active substances) in RxNorm, only 330 (3.6%) appeared in the sample corpus.

We searched the corpus for recognition errors concerning trade names, active ingredients and their dosages by comparing the manual transcriptions to the best paths in the word graphs, and a list of the mismatches (i.e. recognition errors) and their frequencies was compiled. It turned out that 39.3% of all trade names and active ingredients were recognized incorrectly. The average ASR word-error rate of the reports was 38.1%. Aproximately 1-2% of the trade names were not covered by RxNorm.

4 Approach

Our approach consists of a generation mechanism which anticipates possible spoken forms for the content of the knowledge base. The word graphs are searched for trade names or active substances and, subsequently, matching dosages. New concept edges are inserted if valid prescriptions are found in the word graph.

4.1 Detecting Medications in the Word Graph

The (multi-edge) word graphs are scanned, and the words associated with each edge are compared to the medications in the knowledge base. Figure 1 shows a word graph consisting of hypotheses generated by ASR, which is the input to our system. The dashed edges indicate the best path, while dotted lines are hypotheses which are not on the best path.

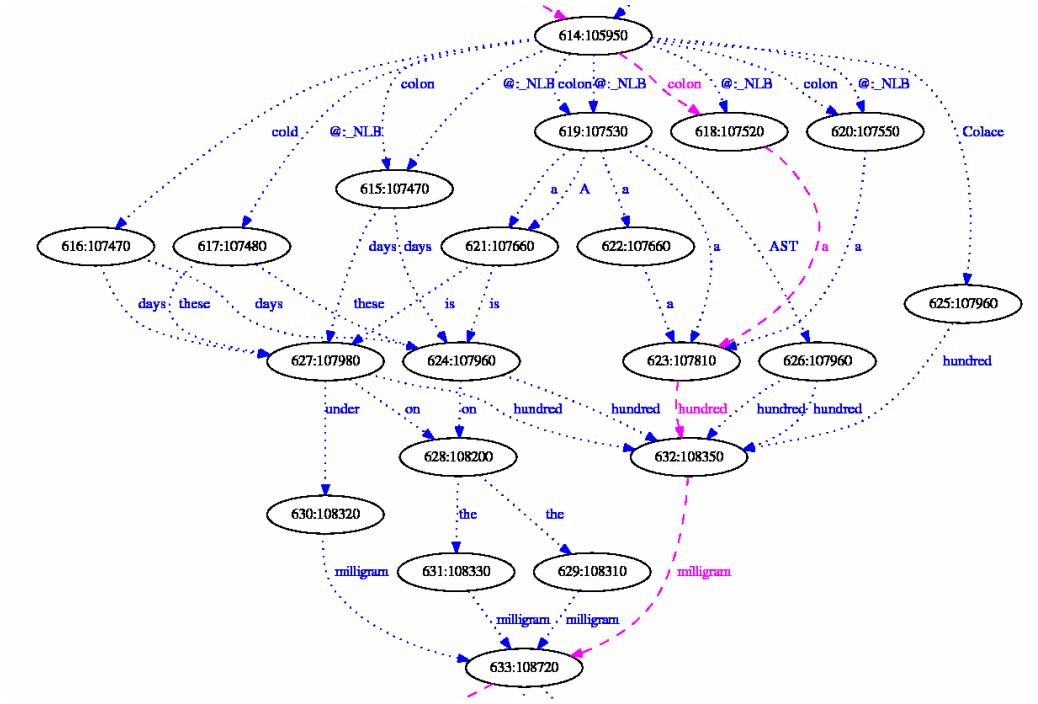

Figure 1: Sample word graph fragment

In case a match, i.e. a trade name or an active substance, is found, all edges succeeding the medication edge are searched for dosage values and dosage units. So far, we only examine the context to the right-hand side; in the data, we did not encounter any medications where the dosage occurred before the trade name or active substance. The following kinds of fillers between the trade name or active substance and the dosage are allowed: 'to' and 'of' as well as non-utterances such as *hesitation*, *noise* and *silence*; in the corpus, we did not encounter any other fillers.

4.2 Generation of Spoken Forms and Mapping

The medication found in the word graph is looked up in RxNorm, and all possible spoken forms of valid dosage values and dosage units for this medication are generated. Spoken forms for the medication names consist of the trade names and the active substances. Variation in the pronunciation of the trade names or active substances is handled by the ASR recognition lexicon. For generating spoken forms of the dosage values, finite-state tools were used. For dosage units, we wrote a small grammar. Looking at two examples, the medication *Synthroid*® and *Colace*® (the latter appears in the word graphs in Figure 2 and Figure 1), the spoken forms shown in Table 1 are generated. Each box contains the alternative spoken variants. *Synthroid*® contains the active substance Thyroxine and *Colace*® contains the active substance Docusate; users may either refer to the trade name or the active substance, so both possibilities are generated for each medication and dosage. RxNorm does not contain the dosage unit 'mcg' (micrograms), which occurred in the reports. Therefore, microgram dosage values were converted to milligrams. Since both 'miligram(s)' and 'microgram(s)' may occur for *Synthroid*®, dosage values for both dosage units are generated. Although strictly, 'twenty five' and 'twenty-five' are identical spoken forms, both versions may appear in the word graph and thus are provided by our system.

Sometimes, a medication may contain several active substances, e.g. *Hyzaar*®, a medication against high blood pressure:

Hydrochlorothiazide 12.5 MG / Losartan 50 MG Oral Tablet [Hyzaar]

trade name/ active substance	dosage value	dosage unit
'Synthroid' 'Thyroxine'	'zero point zero two five' 'zero point O two five' 'O point zero two five' 'O point O two five' 'point zero two five' 'point O two five'	'milligram' 'milligrams'
	'twenty five' 'twenty-five' 'two five'	'microgram' 'micrograms'
'Colace' 'Docusate'	'one hundred' 'a hundred' 'hundred'	'miligram' 'miligrams'

Table 1: Generated spoken forms found in the word graph

In these cases, the generation of possible spoken forms also includes different permutations of substances, as well as a spoken forms containing the dosage unit either only at the end or after each value if the dosage unit is identical.

4.3 Inserting Concept Edges

The sequences of words which constitute the word graph are compared to the spoken forms generated for the RxNorm knowledge base. The active substances or trade names serve as a starting point: in case a trade name is found in the word graph, the spoken forms for dosages of all active substances are generated in all permutations. If an active substance is found in the word graph, only the spoken forms for the substance dosage are searched in the word graph.

A new concept edge is inserted into the word graph for each path matching one of the generated spoken forms of the medications data base. The inserted concept edges span from the first matching node to the last matching node on the path. Figure 2 shows the word graph from Figure 1 with an inserted concept edge (in bold). For each inserted concept edge, new concept-edge attributes are assigned containing the IDs of the original edges as children, their added scores plus an additional concept score and the sequence of words. Since no large-scale experiments have yet been carried out, so far the concept score which is added to the individual scores of the children is an arbitrary number which improves the score of the medication subpath in constrast to paths which do not contain valid medication information. If several competing medication paths are found, a concept edge is inserted for each path, and the concept edges can be ranked according to their acoustic and language-model scores.

5 Evaluation

In the first step, we examined a report sample in order to determine if there are cases where a valid prescription is recognised although the physician did not mention a prescription. We did not encounter this phenomenon in our report corpus.

We then applied our method to a sample of 924 word graphs. In this sample,

- 481 valid dosages could be found, although
- only 325 of these were on the best path.

With our approach, for the 156 prescriptions (32%) which were not on the best path, alternatives could be reconstructed from the word graph. Based on the inserted concept edges, the best path can be rescored.

In order to measure recall, i.e. how many of all existing prescriptions in the reports can be detected with our knowledge base, we manually checked a sample of 132 reports (containing manual transcriptions and ASR results). In this sample, 85 errors concerning medications and/or prescriptions occurred. For 19 of the 85 errors, the correct result was contained in the word graph. For 8 errors, it could be reconstructed. So about 9% of the errors concerning medications can be corrected in our sample. For the cases where the prescription could not be reconstructed although it was contained in the word graph, an analysis of the errors is shown in Table 2.

Since new medications are constantly being released, and trade names change frequently, mismatches may be due to the fact that our version of RxNorm was from a more recent point in time than the report corpus. We assume that under real-world conditions, both RxNorm and the medications prescribed by physicians reflect the current situation.

Some problems concerning medication names and dosage units were caused by missing spoken forms containing abbreviations, e.g. of dosage units (*mg* vs. *mg/ml*) or names (*Lantus* vs. *Lantus insulin*). Here, the coverage needs to be improved.

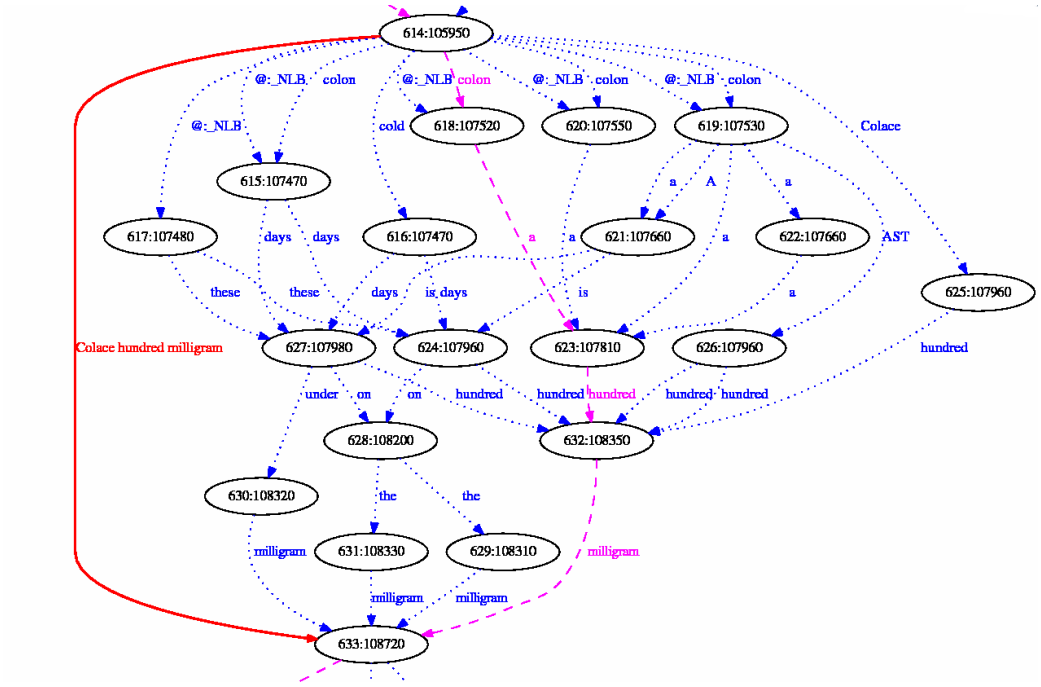

Figure 2: Sample word graph fragment with inserted concept node (left)

Table 2: Error types found in manual evaluation

type of error	**#**	**example**	
		Word Graph	**RxNorm**
differences in medication names between the knowledge base and the word graph	3	*Cardizem CD 120 mg*	*Cardizem 120 mg*
differences in dosage values between the knowledge base and the word graph	4	*Tapazole 60 mg*	*Tapazole 10 mg*
differences in dosage units between the knowledge base and the word graph	4	*Epogen 20000 units*	*Epogen 20000 ml*

There are also cases where two medications appear in the word graph, and both had the valid prescription strength, therefore the system was not able to determine the correct medication.

6 Conclusion

In this paper, we present an attempt to reduce the number of speech-recognition errors concerning prescriptions of medications based on a domain-specific knowledge base. Our approach uses word graphs as input and creates new versions of the word graph with inserted concept edges if more plausible prescriptions are found. The concept edges can be used for rescoring the best path. An evaluation showed that 32% of prescriptions found in the word graphs were not on the best path but could be reconstructed. The manual evaluation of 132 reports shows that our method covers 42% of the prescriptions which are actually spoken during the dictation.

At present, we have only investigated the reduction of medication misrecognitions in our evaluation. In a larger evaluation, we will determine the actual impact of our method on the word-error rate of medical reports. Furthermore, we are working on integrating additional available knowledge sources so that the plausibility of prescriptions can also be as-

sessed from a broader medical point of view, e.g. in case two subsequent prescriptions are encountered in the word graph which are incompatible due to drug interactions. As a next step, the system can be extended to compare the prescriptions with the patient record, e.g. if a patient has medication allergies. So far, our simple solution integrating only available, constantly updated knowledge about medications has already turned out to be a good starting point for rescoring word graphs based on domain knowledge.

Acknowledgments

The work presented here has been carried out in the context of the Austrian KNet competence network COAST. We gratefully acknowledge funding by the Austrian Federal Ministry of Economics and Labour, and ZIT Zentrum fuer Innovation und Technologie, Vienna. The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Transport, Innovation, and Technology and by the Austrian Federal Ministry for Science and Research. The authors would like to thank the anonymous reviewers for their helpful comments.

References

A. Alapetite, A., H.B. Andersen, H.B. and M. Hertzumb. Acceptance of speech recognition by physicians: A survey of expectations, experiences, and social influence. *International Journal of Human-Computer Studies* **67**(1) (2009) 36–49

D. Bühler, W. Minker and A. Elciyanti. Using language modelling to integrate speech recognition with a flat semantic analysis. In: *6th SIGdial Workshop on Discourse and Dialogue*, Lisbon, Portugal (September 2005) http://www.sigdial.org/workshops/workshop6/proceedings/pdf/86-paper.pdf.

S. Gold, N. Elhadad, X. Zhu, J.J. Cimino, G. Hripcsak. Extracting Structured Medication Event Information from Discharge Summaries. In: *Proceedings of the AMIA 2008 Symposium*.

I. Gurevych and R. Porzel. Using knowledge-based scores for identifying best speech recognition hypothesis. In: *Proceedings of ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, Chateau-d'Oex-Vaud, Switzerland (2003) 77–81 http://proffs.tk.informatik.tu-darmstadt.de/TK/abstracts.php3?lang=en&bibtex=1&-paperID=431.

R. Porzel, I. Gurevych and C. Müller. *Ontology-based contextual coherence scoring*. Technical report, European Media Laboratory, Heidelberg, Germany (2003) http://citeseer.ist.psu.edu/649012.html.

S.R. Halgrim, F. Xia, I. Solti, E. Cadag and O. Uzuner. Statistical Extraction of Medication Information from Clinical Records. In: *Proc. of AMIA Summit on Translational Bioinformatics*, San Francisco, CA, March 10-12, 2010.

D.A. Lindberg, B.L. Humphreys and A.T. McCray. The unified medical language system. *Methods of Information in Medicine* **32**(4) (August 1993) 281–291 http://www.nlm.nih.gov/research/umls/.

S. Liu, W. Ma, R. Moore, V. Ganesan and S. Nelson. Rxnorm: Prescription for electronic drug information exchange. *IT Professional* **7**(5) (September/October 2005) 17–23

K. Macherey, O. Bender and H. Ney. Multi-level error handling for tree based dialogue course management. In: *Proceedings of ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, Chateau-d'Oex-Vaud, Switzerland (2003) 123–128, http://www-i6.informatik.rwth-aachen.de/~bender/papers/isca_tutorial_2003.pdf.

M. Oerder and H. Ney. Word graphs: An efficient interface between continuous speech recognition and language understanding. In: *Proc. IEEE ICASSP'93*. Volume 2. 119–122.

J. Patrick and M. Li. A Cascade Approach to Extracting Medication Events. In: *Proc. Australasian Language Technology Workshop (ALTA) 2009*.

R. Sarikaya, Y. Gao and M. Picheny. Word level confidence measurement using semantic features. In: *Proc. of IEEE ICASSP2003*. Volume 1. (April 2003) 604–607.

S. Seneff and J. Polifroni. Dialogue Management in the MERCURY Flight Reservation System. In: *Satellite Dialogue Workshop, ANLP-NAACL*, Seattle (April 2000).

K.D. Voll. *A Methodology of Error Detection: Improving Speech Recognition in Radiology*. PhD thesis, Simon Fraser University (2006) http://ir.lib.sfu.ca/handle/1892/2734.

K. Wang, Y.Y. Wang and A. Acero. Use and acquisition of semantic language model. In: *HLT-NAACL*. (2004) http://www.aclweb.org/anthology-new/N/N04/N04-3011.pdf.

R. Zhang and A.I. Rudnicky. Word level confidence annotation using combinations of features. In: *Proceedings of Eurospeech*. (2001) http://www.speech.cs.cmu.edu/Communicator/papers/-RecoConf2001.pdf.