Österreichisches Forschungsinstitut für / Austrian Research Institute for / Artificial Intelligence

TR-2009-05

Stephanie Schreitter, Alexandra Klein,
Johannes Matiasek, Harald Trost

Using Domain Knowledge to Improve Automatic Speech Recognition: Correcting Errors in Prescriptions of Medications

• Freyung 6/6 • A-1010 Vienna • Austria •
• Phone: +43-1-5336112 •
• mailto:sec@ofai.at •
• http://www.ofai.at/ •

Österreichisches Forschungsinstitut für / Austrian Research Institute for / Artificial Intelligence

TR-2009-05

Stephanie Schreitter, Alexandra Klein, Johannes Matiasek, Harald Trost

Using Domain Knowledge to Improve Automatic Speech Recognition: Correcting Errors in Prescriptions of Medications

The Austrian Research Institute for Artificial Intelligence is supported by the Federal Ministry of Education, Science and Culture.

Citation: Stephanie Schreitter, Alexandra Klein, Johannes Matiasek, Harald Trost: Using Domain Knowledge to Improve Automatic Speech Recognition: Correcting Errors in Prescriptions of Medications. Technical Report, Österreichisches Forschungsinstitut für Artificial Intelligence, Wien, TR-2009-05.

Using Domain Knowledge to Improve Automatic Speech Recognition: Correcting Errors in Prescriptions of Medications

Stephanie Schreitter[1], Alexandra Klein[1], Johannes Matiasek[1], and Harald Trost[2]

[1] Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
[2] Department of Medical Cybernetics and Artificial Intelligence, Center for Brain Research, Medical University of Vienna, Austria

Abstract. We present an approach to improving automatic speech recognition (ASR) for the creation of medical reports by analyzing hypotheses in the word graph based on background knowledge. Our application area is prescriptions of medications, which are a frequent source of misrecognitions: In a sample of 123 reports, we found that no less than about a third 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 – a significant potential for improvement. To realize this potential, we have built a knowledge base of medications based on information contained in the Unified Medical Language System (UMLS). This knowledge base contains trade names, active substances, strengths and dosages. Based on this representation, we generate a variety of linguistic realizations for prescriptions. Whenever an inconsistency in a prescription is encountered in 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 about 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 professional applications, namely in the domain of medical reporting where there is an ever growing demand for documentation. Users appreciate the fact that speech-recognition provides a hands-free input mode, which is important as they often simultaneously handle documents such as notes and X-rays, and that the records can be accessed immediately after their creation [1]. 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.

Since medical texts largely belong to limited domains, many recognition errors could be avoided by consulting background knowledge concerning the respective domains. In general, modelling domain knowledge is often difficult as i) the developers of knowledge bases usually lack expert knowledge, ii) the knowledge bases are tailored to specific applications and cannot be reused, and most of all, iii) the creation is very costly. Fortunately, in the medical domain, many knowledge sources are freely available.

We have considered prescriptions of medications a good starting point as they are common and frequent in the various medical fields. Furthermore, from a cursory glance at recognition results, we had the impression that many recognition errors occur in trade names and dosages. This comes a no suprise as in all domains, proper names and digits, i.e. trade names or active substances and dosage values in the medication domain, are frequently misrecognized.

For our approach, we have extracted and adapted information about medications from the Unified Medical Language System (UMLS)[3] [2]. 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 a first step, we assessed the potential for improvement by analyzing medical reports. It turned out that in a sample of 123 dictated medical reports which were processed by a speech-recognition system, no less than about a third of the active substances or trade names and dosages were recognized incorrectly. Examining the word graphs of the reports, we realized that in about 25% 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 for detecting speech-recognition errors and semantic rescoring of word-graph hypotheses. Then, we will describe our approach for adapting information about medications from the UMLS to enhance the word graph with concept nodes representing domain-specific information. Finally, we will illustrate the potential of our approach by means of a manual, qualitative evaluation of word graphs for medical reports which were processed by our system.

2 Error Handling and Semantic Rescoring

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. Language-model scores and acoustic scores are combined in the creation of word graphs for the user utterance. It has often been assumed that the introduction of an explicit representation of the utterance meaning will improve recognition results. In fact, this works best in limited domains: the larger an

[3] http://www.nlm.nih.gov/research/umls/

application domain, the more difficult it is to build an optimal knowledge representation which anticipates 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. [3]).

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. For example, error handling introduces a post-processing step (after speech recognition has handled the user input) to support the dialogue manager in dealing with inconsistencies [4]. As far as dictated input is concerned, which is not controlled by a dialogue manager, [5] 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 [6, 7]. [8, 9] presents 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 ([10, 11]) 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 have been specially created for the applications or 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.

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, consisting of manual transcriptions and word graphs created by a speech-recognition system.

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 [12]. 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 at 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, which does not occur in our report sample, is oral tablet.

Our version of RxNorm contains 4,489 active substances and 6,310 trade names (7,597 trade names counting the different dosages). The active ingredients in RxNorm are associated with Anatomical Therapeutic Chemical (ATC) Codes.

3.2 Sample Corpus

The sample corpus consists of 123 medical reports created by physicians from various medical fields and hospitals. For each report, the following data sources were available:

- word graphs with marked best paths, representing the recognition result,
- manual transcriptions, indicating what was actually said during dictation.

We searched the corpus for recognition errors concerning trade names, active ingredients and their dosages by comparing the manual transcriptions to the best path in the word graph, and we compiled a list of the mismatches (i.e. recognition errors) and their frequencies. It turned out that 31.7% of all trade names and active ingredients were recognized incorrectly.

In 64% of these cases, a medication was actually spoken but misrecognized, while in 36% of the cases, a medication was recognized although this part of the user utterance did not contain a medication. 48% of the dosage errors are combined with a misrecognition of the active substance or trade name.

In the report sample, approximately 1-2% of the trade names are not contained in RxNorm. These cases had to be identified manually as the medications could not be found using the RxNorm list. 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.

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.

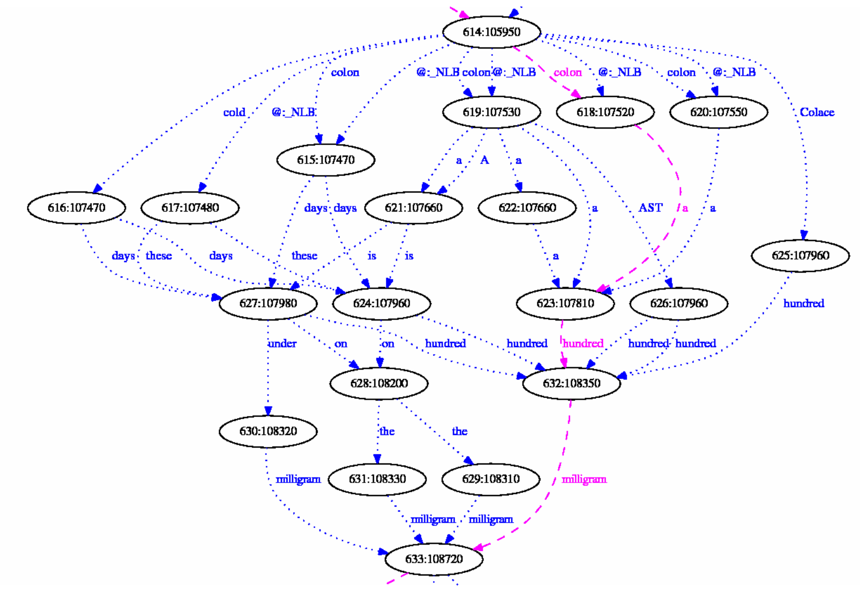

Fig. 1. Sample word graph fragment

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*.

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 medications include the trade names and the active substances. 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 possibilies 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.

Table 1. Generated spoken forms to be found in the word graph

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'

Sometimes, a trade name contains several active substances, e.g. *Hyzaar*®, a medication against high blood pressure:

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

In these cases, the generation of possible spoken forms also includes different permutations of the two substances, as well as a spoken forms containing the dosage unit either only at the end or after both values if the dosage unit is identical. Up to four active substances may be listed per trade name and can be handled this way.

4.3 Inserting Concept Edges

The sequences of words which constitute the word graph are compared to the spoken forms generated for RxNorm. The active substance or trade name serves 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.

For rescoring, new concept edges are inserted into the word graph representing 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 one of the path. Figure 2 shows the word graph from Figure 1 with an inserted concept edge (in bold).

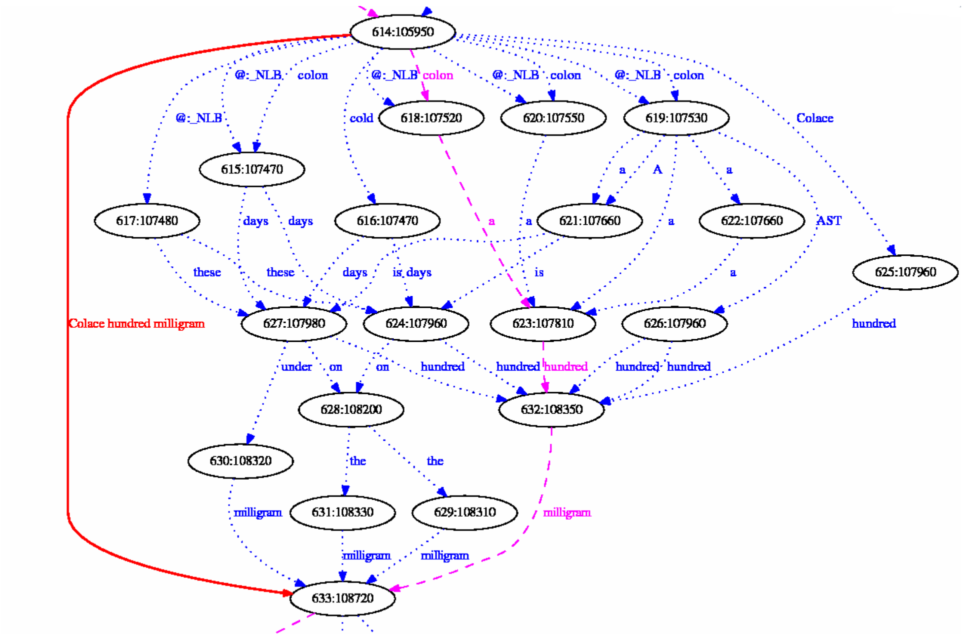

Fig. 2. Sample word graph fragment with inserted concept node (left)

For each concept edge, new concept-edge attributes are assigned containing the IDs of the original edges as children edges, their added scores plus an additional concept score and the sequence of words. Thus, the new concept edges complement the edges which have been computed based on the acoustic and language-model scores and can be used for rescoring the word graph.

5 Evaluation

We applied our method to a report sample of 123 reports. In this sample, 85 errors concerning medications and/or dosages occurred. In 19 out of the 85 cases, the correct prescription was contained in the word graph, but was not on the best path. With our approach, 8 of these 19 prescriptions could be reconstructed from the word graph. Based on the inserted concept edges, the best path can be rescored.

For the remaining 11 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.

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 dosage units between the knowledge base and the word graph	4	*Epogen 20000 units*	*Epogen 20000 ml*

Some problems concerning medication names and dosage units were caused by missing spoken forms containing abbreviations, e.g. abbreviated dosage units (*mg* vs. *mg/ml*) or names (*Lantus* vs. *Lantus insulin*). Here, it is easy to improve the system coverage. There are also cases where two medications appeared in the word graph, and both had the valid prescription strength 10 mg, therefore the system was not able to determine the correct medication.

6 Conclusion

In this paper, we have presented an attempt to reduce the number of speech-recognition errors concerning prescriptions of medications based on a domain-specific knowledge base. An evaluation showed that for misrecognitions of trade names or active substances combined with dosages, about 10% of the errors can be avoided. Among the errors where the correct prescription was contained in the word graph, about half of the misrecognitions could be prevented.

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. From the valid expressions, a grammar could easily be derived which can be used as a third source of information, in addition to the acoustic scores and the language-model score, during speech recognition to determine the best path in just one step.

At present, we have only investigated the reduction of medication misrecognition 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 plan to integrate additional available knowledge sources so that the plausibility of prescriptions can also be assessed 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 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.

References

1. Alapetite, A., Andersen, H.B., Hertzumb, M.: Acceptance of speech recognition by physicians: A survey of expectations, experiences, and social influence. Inernational Journal of Human-Computer Studies **67**(1) (2009) 36–49
2. Lindberg, D.A., Humphreys, B.L., McCray, A.T.: The unified medical language system. Methods of Information in Medicine **32**(4) (August 1993) 281–291 http://www.nlm.nih.gov/research/umls/.
3. Seneff, S., Polifroni, J.: Dialogue management in the mercury flight reservation system. In: Satellite Dialogue Workshop, ANLP-NAACL, Seattle (April 2000)
4. Macherey, K., Bender, O., Ney, H.: 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.

5. Voll, K.D.: A Methodology of Error Detection: Improving Speech Recognition in Radiology. PhD thesis, Simon Fraser University (2006) http://ir.lib.sfu.ca/handle/1892/2734.
6. Zhang, R., Rudnicky, A.I.: Word level confidence annotation using combinations of features. In: Proceedings of Eurospeech. (2001) http://www.speech.cs.cmu.edu/Communicator/papers/RecoConf2001.pdf.
7. Sarikaya, R., Gao, Y., Picheny, M.: Word level confidence measurement using semantic features. In: Proc. of IEEE ICASSP2003. Volume 1. (April 2003) 604–607
8. Gurevych, I., Porzel, R.: 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.
9. Porzel, R., Gurevych, I., Mueller, C.: Ontology-based contextual coherence scoring. Technical report, European Media Laboratory, Heidelberg, Germany (2003) http://citeseer.ist.psu.edu/649012.html.
10. Wang, K., Wang, Y.Y., Acero, A.: Use and acquisition of semantic language model. In: HLT-NAACL. (2004) http://www.aclweb.org/anthology-new/N/N04/N04-3011.pdf.
11. Bühler, d., Minker, W., Elciyanti, A.: 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.
12. Liu, S., Ma, W., Moore, R., Ganesan, V., Nelson, S.: Rxnorm: Prescription for electronic drug information exchange. IT Professional **7**(5) (September/October 2005) 17–23