Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)

doi:10.3233/978-1-61499-133-5-201

Knowledge Acquisition Tool for Dialogue Systems

Raul SIREL[a,1]
[a] *University of Tartu*

Abstract. Knowledge base construction for dialogue systems is a time-consuming process which requires considerable amount of attention from engineers. In this paper a tool for acquiring knowledge for dialogue systems is presented. The tool has been developed to facilitate the knowledge base creation process and utilizes adjacency pairs (or Frequently Asked Questions pairs) as an input. The tool uses three simplistic text-mining algorithms for finding keywords from input pairs and outputs keywords-answer pairs to be used by knowledge engineer or dialogue system developer in the process of knowledge base construction.

Keywords. dialogue systems, knowledge acquisition, Estonian language

Introduction

Knowledge acquisition and representation is a critical task in the development process of intelligent computer systems such as natural language dialogue systems. Androutsopoulos and Aretoulaki [1] describe them as natural language interaction systems; that is, systems that allow users to formulate their requests in natural language. Though, according to Jurafsky and Martin [2] dialogue systems most often communicate through speech rather than text (because in mobile situations input devices such as keyboards are not available), the value of text-based dialogue systems should not be underestimated.

The most common use of written dialogue systems today is answering questions delivered in natural language [3]. Numerous dialogue agents exist that commune in natural Estonian such as Theatre Agent [4] and Tooth Fairy [5] which are both implimented using a domain-independent dialogue system framework [6]. Despite the multitude of dialogue agents created with the described framework, the knowledge base construction for those systems has proven to be extremely time-consuming task, which has motivated the creation of the tool described in this article.

It is the purpose of this paper to present the text-mining based tool created for facilitating the knowledge acquisition process for dialogue systems. The article also

[1] Corresponding Author: PhD student, University of Tartu, Ülikooli 18, 51014 Tartu, Estonia; E-mail: rsirel@ut.ee

proposes possible target resources to be used for acquiring knowledge using this tool and gives an overview where it may be (and already has been) utilised.

The tool was created to be used for Estonian, but the concept is appicable for other languages as well. The tool was created using Python programming language.

1. The Tool

The tool uses adjacency pairs as input. An adjacency pair is a unit of conversation that contains an exchange of one turn each by two speakers [7]. Such input may be collected from annotated dialogue corpora (such as Estonian Dialogue Corpus [8]) or more easily from Frequently Asked Questions' pages on the web. The FAQs are a genre of internet texts that explain somewhat trivial questions regarding some limited topic, they also may be considered structurally very similar to adjacency pairs since they consist of two turns – question and answer.

The tool collects relevant keywords from adjacency pairs provided by the user and outputs them as keywords-phrase pairs to be used in the knowledge base construction. It is assumed that such keywords-phrase pairs would considerably facilitate the knowledge base construction process.

The tool implements 5 simple text-mining algorithms and uses morphological analyser, morphological disambiguator and Estonian Wordnet as external resources. The described tool has already been utilised in knowledge base construction for various dialogue agents, for example a system for preliminary periodontal consulting [9].

1.1. Morphological analyser and disambiguator

Estonian is morphologically quite a complex language [10] and requires additional software to successfully apply text-mining to it. Lemmatisation software is usually utilised to reduce the size of the lexicon. For example words *poiss (boy)* and *poisid (boys)* are reduced to their lemmas which is in this case is *poiss (boy)*. But quite often the process of lemmatisation faces the problem of homonymy in which case the software is not able to choose the correct lemma simply by analysing the word. For example the word *lood* has two different lemmas: *lood (builder's level)* and *lugu (story, tale)*. In situations like this, morphological disambiguation is required to solve the problem: disambiguator analyses the local context (surrounding words) to choose the correct lemma.

Morphological Disambiguator implemented in the tests described in this paper was developed by Filosoft Ltd [11] and uses hidden Markov model.

1.2. Wordnet

While applying text-mining to natural language, it is crucial to take into consideration that synonymy allows us to mark the same object with different markers. Wordnet is a type of lexical database where concepts are not organised alphabetically, but by semantic relations [12]. In this paper Estonian Wordnet [13] is used to find synonyms and hyperonyms to keywords mined from the FAQ sets.

2. Algorithms

The first algorithm is based on a naïve hypothesis that more frequent lemmas in the input sets are more suitable candidates for keywords. The first algorithm consists of three basic steps (see Figure 1).

```
1. Lemmatise the input FAQ set
2. Find most frequent lemmas
3. Find synonyms and hyperonyms for the lemmas found
```

Figure 1. First algorithm

The second algorithm is based on the set theory, from which the concept of intersection is introduced to mine for keywords. The main hypothesis is that more important lemmas (more suitable for keywords) exist both in question and answer (see Figure 2).

```
1. Lemmatise the question and answer to separate sets
2. Find the intersection of the sets
3. Find synonyms and hyperonyms to each lemma in the intersection
```

Figure 2. Second algorithm

The third algorithm is similar to the second one; this too is based on finding the intersection of the lemma sets. Additionally a filtering mechanism to separate the noise from productive keywords is introduced (see Figure 3).

```
1. Lemmatise the adjectives, nouns and verbs found in the question
   and answer to separate sets
2. Find the intersection of the sets
3. Remove stop words from the intersection
4. Find synonyms and hyperonyms to each lemma in the intersection
```

Figure 3. Third algorithm

3. Evaluation

For the quality and accuracy of generated keywords is somewhat subjective to the knowledge engineer, in addition to the subjective qualitative approach, a simplistic quantitative approach has been used to evaluate the system's output.

Table 1. Numerical results of the algorithms

	Algorithm 1	Algorithm 2	Algorithm 3
Number of sets used	100	100	100
Number of keywords found	383	332	287
Number of sets that produced minimum of 1 keyword	100	90	81
Average number of keywords per set	3,83	3,69	3,54

Table 1 shows that the third algorithm produced the smallest number of keywords (followed by the second algorithm) which is a direct result of the filtering methods used. Though only 81% (90% with the second algorithm) of the initial sets produced at least 1 keyword, the qualitative analysis of the results clearly demonstrated that the keywords found by the 3rd algorithm were the most accurate (followed by the 2nd algorithm) and required very little post-mortem filtering.

Most of the incorrect keywords from algorithm 2 and 3 were produced by finding synonyms and hyperonyms from Wordnet (which can be disabled from the tool's interface). However, in many cases Wordnet had rather positive effect – many of the suggested synonyms and hyperonyms appearted to be non-trivial and helpful in the knowledge base construction process.

Conclusion

Knowledge acquisition from free texts is a significant goal in the fields of text mining and knowledge engineering, and even though substantial research has been conducted, further investigation is still required.

Current paper focused on acquiring knowledge from adjacency pairs (and Freqently Asked Questions' sets) by applying three different text-mining algorithms. A tool was created to automatically find keyword-phrase pairs from aforenamed resources. The results showed that it is possible to facilitate the knowledge base creation process even by using extremely simplistic and naïve methods.

Despite the fact that described methods were developed to be used for Estonian, their simplicity makes them rather language independent. For morphologically complex languages, morphological analyser (and disambiguator) should be adapted for input lemmatisation. Also lexical resources such as Wordnet or some other thesauri should be used for finding synonyms (and hyperonyms).

Acknowledgements

This work is supported by the European Regional Development Fund through the Estonian Centre of Excellence in Computer Science (EXCS), the Estonian Science Foundation (grant 9124), and the Estonian Ministry of Education and Research (projects SF0180078s08, and EKT11005).

References

[1] Androutsopoulos, I; Aretoulaki, M. *Natural Language Interaction.* The Oxford handbook of computational linguistics. Oxford University Press, 2003.
[2] Jurafsky, D; Martin, J. H. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.* Pearson/Prentice Hall, 2009.
[3] Treumuth, M. *A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects*. PhD Thesis. Tartu University Press, 2011.
[4] Teatriagent. http://www.dialoogid.ee/teatriagent/. Checked 18.06.2012.
[5] Hambahaldjas. http://www.dialoogid.ee/hambahaldjas/. Checked 18.06.2012.
[6] Treumuth, M. A Framework for Asynchronous Dialogue Systems. *Frontiers in Artificial Intelligence and Applications: Human Language Technologies — The Baltic Perspective*. IOS Press, 2010.
[7] Levinson, S.C. *Pragmatics.* Cambridge University Press, 1983.
[8] Estonian Dialogue Corpus (EDiC). http://math.ut.ee/~koit/Dialoog/EDiC.html. Checked 18.06.2012.
[9] Pähkla, E.R.; Treumuth, M. A computer program answering questions in preliminary periodontal consulting. *Journal of Clinical Periodontology*, Volume 39, Issue Supplement s13, 2012.
[10] Erelt, M. *Estonian Language.* Linguistica Uralica Supplementary Series vol 1. Estonian Academy Publishers, 2003.
[11] Filosoft Ltd. http://www.filosoft.ee. Checked 18.06.2012.
[12] Miller, G. A; Beckwith, R; Fellbaum, C. D; Gross, D; Miller, K. J. Introduction to WordNet: An On-line Lexical Database. *Practical Lexicography*, 2008.
[13] TEKsaurus. http://www.cl.ut.ee/ressursid/teksaurus/. Checked 18.06.2012.