ReMa L&C, Introduction to Language and Speech Technology

Final Exam, 2nd Chance, April 5, 2018

This exam consists of 3 questions. Every question is worth 10 points. Subquestions are marked for their points.

First read all questions carefully and then choose in which order you will answer them. You can put the answers to the questions (and even parts of questions) in any order you want, as long as you clearly mark which answers belong to which questions. Also put your name and student number on any handed in sheet of paper and/or computer file.

Please note that answers should be supported by argumentation. However, restrict yourself to relevant arguments and do not just try to write down everything you know. The presence of excessive irrelevant information may lead to a deduction of points.

Background information for the questions: Language adaptation.

Building a full range of NLP systems for a specific language is quite a lot of work. One may wonder to which degree it would be possible to replace systems specifically made for the language itself by systems made for English, in combination with a statistical machine translation system. In this exam, we will explore this idea. The system we would like to build should be able to carry on a conversation with the user and provide various kinds of information that the user asks for, most likely information that can be found on internet. The user would be typing in his/her own language, the system would do its work in English, and a statistical machine translation system would translate user statements into English, and finally the system translates system responses from English back into the user's language.

1) Let us start with dialogue-based interface.
    a) Describe the main differences you expect between i) dialogues between native speakers of English and ii) dialogues between a native speaker of English and a native speaker of another language, supported by a (not fully capable, either human or machine) translator. Use the concepts that are given in Section 24.1. (4 points)
    b) Describe how you would adapt the various parts of the dialogue component of your system, so that it can cope as well as possible with these differences. (3 points)
    c) Describe how well you expect the complete system to work. Explain your expectations. (2 points)

2) Then we proceed to the information gathering component(s). Assume that the user is satisfied enough with the dialogue interface, and that he/she will use full sentences for asking questions rather than a series of query words. Also assume for this question that a significant amount of the desired information is only present on internet in the user's language, and that the subquestions below are about finding that information.

   a) Explain how you would use your software to find relevant documents for answering queries. Discuss two cases, and also compare the two:

      i. Use your existing document retrieval (developed for English) on user language queries and documents. This means that there is no translation involved whatsoever. You just use the query terms in the user's language to search for documents in the user's language, even though your retrieval system is specifically made to search in English)

      ii. Use English language retrieval on automatically translated queries and automatically translated documents. In this case we have to assume that we search a data collection rather than the full interne).

      (4 points)

   b) For extracting the relevant knowledge from the retrieved documents, you will have to use your English knowledge extraction components. Describe in which components you expect problems caused by having to work on translated material (i.e. bigger problems than you would already have when working on English material). Refer to the appropriate sections in the book. (4 points)

   c) Describe to which degree the quality of the knowledge extraction depends on the user's language and the kind of topics that are being discussed. (2 points)

3) Finally, we turn to desired information that is of an international nature and for which many different languages need to be processed. One of the problems that we encounter is that many proper names are spelled differently in different languages, e.g. because of different transliteration strategies.

   As an example of spelling variants, take the name *Muhammad*, the variation of which can be seen in the corresponding Wikipedia page (https://en.wikipedia.org/wiki/Muhammad_(name)). Note that this is just an example. The variation is quite extreme. However, this does not imply that other names will always have less variation. In fact, there are even other effects possible than the ones visible with *Muhammad*, when other source and target languages come into play.

   Note that the task is not to find an internationally accepted normal form for a given name, but rather to compare a string (that is assumed to be a name) with a list of potential names and find the name in the list which is most likely to be meant by the given string.

   a) Discuss the relative usability of the Levenshtein and Viterbi algorithms for normalizing spelling variants of names, for the purpose of doing information retrieval in each document language. Describe the advantages and disadvantages of both algorithms. (4 points)

   b) Discuss other machine learning options for spelling normalization of names, including which features you would give the learners access to. (4 points)

   c) Discuss the use of statistical machine translation for spelling normalization of names. (2 points)