Naam & Studentnummer

Language Technology
Fall 2009
Docent: John Nerbonne
Exam — 90 min.
9:00 am — 26 Oct. 2009

Important Instructions

1. Write your name and student number on this exam, and write your initials on *all* answer sheets. This minimizes the chance that anything is lost.

2. No use of books or study materials is allowed.

3. You have 90 min. to complete this (language) half of the exam (and another 90 for the speech exam).

4. If you find one or another question problematic, try to explain why as explicitly as possible.

5. The language exam will be graded separately from the speech exam, so please keep the answer sheets separate!

6. One question (on corpora) counts 10 points, the others all 5 points each. There are 50 points in total on the language half of the exam.

1 Basic ideas

1. Consider an application of language technology in which geographical expressions such as *Groningen, London, Dutch capital city, Mississippi, or northern Uganda* must be identified and "geo-referenced", in other words, where the application must (automatically) say what part of the world the expressions refers to, perhaps by providing (sets of) longitude-latitude coordinates or perhaps by showing the location on a map. Your task is to check the quality of the technology for an application in which the system is to be incorporated into a browser so that newspaper readers can ask where someplace as they read an article. You work for the newspaper.

1. Explain the difference between EVALUATION and ASSESSMENT on the basis of this application. What is your primary concern as an employee of the newspaper? (**5 pt.**)

   Evaluation aims to measure the generic quality of a language technology component while assessment aims to measure its suitability for a specific task, in this case how well place names are recognized and geo-referenced. This might also be the subject of evaluation, but since assessment is specific it will test a specific newspaper, perhaps with a disproportionate number of local place names.

   Some of you mentioned usability, which is another sensible property to expect of applications, but its not the same thing.

2. Explain how the concepts PRECISION and RECALL might be applied to the job of checking the quality of of the language technology. (Do not merely repeat the general definitions, but say how they should be applied to this application.) (**5 pt.**)

   Precision is the fraction of identifications that are correct, and recall the fraction of geographical expressions in the texts that are correctly identified (or 1-(fraction overseen)).

2 Corpora (10 pt.)

3. Before about 2000 many theoretical linguists collected data by checking with native speakers whether given sentences (or words, or phrases) were "acceptable". They would ask whether examples such as the following were "well-formed":

   > Wat denk je wat hij vroeg wat ik wilde doen?

   There were naturally worries, e.g., that researchers might influence their native speaker "informants" so extensively that they would provide the answers the researchers want. In particular there were many complaints that very unlikely sentences and phrases were reported as acceptable to native speakers. But corpora were too small to expect to find much, and checking the corpora involved too much time and special expertise.

   As corpora have become available researchers have definitely appreciated their value, but most reports do not exactly confirm the earlier worry. Bresnan's article was explicit, however, about another sort of problem. Identify the problem Bresnan focuses on, and explain how corpora have exposed it as another problem of the methodology of interviewing native speakers.

   An ambitious answer might go on to sketch Bresnan's conjecture about the relation between acceptability judgments and corpus frequency and how she has tried to test this conjecture.

Bresnan's article shows that some well-studied structures reported as ill-formed in fact do occur in corpora, which is opposite of the worry that some ill-formed expression might be judged acceptable. Researchers appear not to consider a wide enough range of contexts so that the effect of context on well-formedness is underestimated. Bresnan conjectures that well-formedness judgments may in fact reflect corpus frequencies. She tests this by building a statistical model which reflects corpus frequencies in that it includes factors such as the choice of verb, whether objects are pronominal, and whether they are definite (if non pronominal) She shows that this model predicts the judgments of well-formedness, and incidentally, that some structures which had been reported as ill-formed are actually preferred if the right combination of factors is present.

3 Information Extraction

4. Explain what is meant by 'information extraction' (IE) in particular in contrast to 'information retrieval', and discuss three examples of practical applications of IE, explaining why they are commercially interesting. (**5 pt.**)

IR aims at retrieving documents, and IE at retrieving specific items of information. Some IE applications are

- question-answering, e.g. for customer contact or as an addition to technical documentation.
- document enrichment ("wikifying") for educational purposes
- "clippings services" for companies following their own public relations, or who are interested in specific developments (e.g., markets, competitors, or resources)
- summarization of technical and policy documents
- ontology building (terms and relations)

4 Language Acquisition

5. Techniques from machine learning are widely used in engineering-oriented language technology applications, such as "named-entity recognition", well as in cognitively motivated computational models of human language acquisition. The models of human language acquisition need to meet certain additional criteria (besides being able to learn correctly and effectively) to be plausible models. List and briefly explain two such criteria. (**5 pt**).

   Cognitive requirements that may not be interesting in applications of learning software include:

   - modeling the time path of development and not just the final result
   - learning in away that not only the successful language of the children is modeled, but also the errors
   - incremental processing, i.e. learning as speech is heard or produced
   - learning in spite of a limited capacity for rote memory
   - learning from an amount of data that might realistically be experienced by a child learning a language.
   - learning with limited feedback, at least with respect to issues of form.

6. Give a brief informal definition of the (letter) successor variety algorithm for segmentation. (**5 pt**)

   Given the word list:

   READ, READABLE, READING, READS, RED, ROPE

   what are the successor values after the following strings:
   R
   RE
   READ

   The (letter) successor variety algorithm for segmentation simply counts the number of possible successors at each prefix (initial substring) and hypothesizes that segment divisions are most likely where the variety is high. For the strings above the variety of successors is 2 (R), 2 (RE) and 4 (READ).

5 Machine Translation

7. It has been shown that adding syntactic knowledge to statistical machine translation systems improves performance. Give a brief definition of syntax in your own words and explain why knowledge of syntactic structure could be beneficial in machine translation. (**5 pt.**)

Syntax concerns the combination or words to form phrases and sentences. Because the manners of combination are quite different in different languages, and because meaning – which one tries to preserves in translation – depends on syntax, it is plausible that translation software should improve by incorporating syntax. There are thousands of examples of syntactic differences, but one simple example of a difference between English and Dutch is that Dutch allows one to omit the objects of prepositions more easily. *Daar loopt dhr. Robben, (hij is) de vader van!* It is furthermore clear that syntax can inform disambiguation.

6 Pronunciation Comparison

8. Name the main reason why taking an aggregate view in variationist linguistics is an improvement over investigating only individual features? Explain why computational techniques are needed to measure aggregate differences. (**5 pt.**)

   The linguistic variety spoken at a given settlement is characterized by tens of thousands of features. If we view the geographic distributions of the individual features, we see irregular and imperfect patterns, which, moreover, differ among themselves. Dialectologists examine aggregations of features in order to abstract away from the noise in single features, to avoid the need to choose which features to study, and to obtain a more abstract characterization of the relations among settlements. Computers are needed to deal with the quantity of data involved.

9. Why is a regular cluster map not a good visualization method of aggregate dialect distances? (**5 pt.**)

   A regular cluster map shows a partition of a given language area but does not show whether some partitions are closer than others, nor does it show the relative similarity of points within partitions, all of which are colored the same. It also does not show the certainty with which lines are drawn. If the partition is derived from one of the more popular clustering algorithms, we would like to see how well the number of elements in the partition (the number of areas) is justified, but the map also does not show that.