Natural Language Processing research and development face three main problems: coverage of the lexical and grammatical phenomena of natural languages, robustness in the face of error-prone input, and accuracy of processing. Accuracy is an area of intense research based on publicly available gold-standard resources. Robustness has been addressed by the large-scale adoption of statistical methods, but with limited attention to its systematic evaluation. Coverage, however, seem to have only been addressed in industrial contexts, with less interest from the research community.

Both rule-based and statistics-based NLP systems face the problem of processing input that is different from (i.e., is not covered by) the linguistic phenomena about which a system has information. Unknown terms and unusual sentence structures are the best known problems of coverage. The British National Corpus, for example, uses over 15 million different terms (types), but a standard commercial machine translation dictionary will cover only 300 thousand terms. This mismatch in coverage leads to serious issues of accuracy and usability of machine translation and other NLP systems. To circumvent this problem, common approaches in information retrieval, for example, avoid using any linguistic knowledge at all, and instead focus on statistical techniques that depend on multiple, independent and redundant indicators of content (e.g., re-occurring strings).

More recent approaches (e.g., Callison-Burch, 2007) automatically identify equivalence classes of strings and substrings — i.e., paraphrases — so that a system can choose among the equivalent items the one that it has the most knowledge of. So, for example, a translation system that has a translation for the term troops -- but not for military force, defense, forces, or armed forces -- can translate all of these accurately once it can establish that they are paraphrases of each other, effectively increasing the coverage of the system from one term to five. Successes such as these with lexical paraphrasing (using substitution) opens the door to discussions of the practical usefulness of other kinds of paraphrasing for NLP applications.

The goal of this course is to provide an introduction to the issues and dimensions of paraphrasing and to illustrate their application to text mining. At the end of this course, students will comprehend the issues involved in using paraphrasing to address the general issue of coverage in the applications of their choice. They will also have knowledge of existing techniques and applications to text mining, as well as an awareness of the current challenges that they can help solve.
Tópicos

I. Major problems of NLP (1 hr)

II. Text Mining: Problems and Perspectives (3 hrs)

III. Dimensions and Techniques of Paraphrasing (3 hrs)

IV. Applications of Paraphrasing Technology (3 hrs)
I. Major problems of NLP (1 hr)
   a. Coverage (grammatical, lexical): Paraphrase helps solve this problem
   b. Robustness (errors in input)
   c. Accuracy

II. Text Mining: Problems and Perspectives (3 hrs)
   d. Current approaches focus on words (IR approaches with term vectors)
   e. Sentence structures are starting to be used (eg, Lin, Pasca et al at Google)
   f. Semantic relations can be used as equivalence classes of sentence structures

III. Dimensions and Techniques of Paraphrasing (3 hrs)
   g. Lexical paraphrasing
   h. Syntactic paraphrasing
   i. Collocational paraphrasing
   j. Referential paraphrasing
   k. Semantic and pragmatic issues: defining equivalence classes of sentence components

IV. Applications of Paraphrasing Technology (3 hrs)
   l. Text mining and information extraction
   m. Readability and Content Management
   n. Input normalization for NLP
   o. Machine Translation (eg, Callison-Burch, 2007)

Bibliografía de referencia


Universal Networking Language (http://www.undl.org)