Department of Computer Science החוג למדעי המחשב
University of Haifa אוניברסיטת חיפה

Mount Carmel, 31905 Haifa, Israel Phone: +972-4-8240259 Fax: +972-4-8249331 הר הכרמל, חיפה 31905 טל. 04-8240259 פקס: 04-8249331

Laboratory in Natural Language Processing (203.4650)

Shuly Wintner

Semester A, 2002-3: Wednesday, 16:00–18:00
http://cs.haifa.ac.il/∼shuly/teaching/03/lab/

1 Objectives

The main objective of the Lab is to thoroughly investigate a specific issue in Natural Language Processing by reading research articles and implementing the ideas described in them. This year we will concentrate on approaches to machine learning of natural language morphology. Each team will read, understand, implement and evaluate a particular article describing a specific approach. We will then try to compare the various approaches, and in particular evaluate their applicability to the problem of learning the morphology of Hebrew. An unrelated project deals with compilation of specifications between two existing extended regular languages.

2 List of projects

2.1 A compiler from LEXC and XFST to FSA

Finite-state technology is widely considered to be the appropriate means for describing the phonological and morphological phenomena of natural languages. Several FS "toolboxes" exist which facilitate the stipulation of phonological and morphological rules by extending the language of regular expressions with additional operators. Such toolboxes typically include a language for extended regular expressions and a compiler from regular expressions to finite-state devices (automata and transducers). Unfortunately, there are no standards for the syntax of extended regular expression languages.

The goal of this project is to design and implement a compiler which will translate grammars, expressed in the finite-state toolbox of Xerox (which include two systems, LEXC and XFST), to grammars in the language of the FSA Utils package. For the most part, there is a strong parallelism between the languages, but certain constructs will be harder to translate and will require more innovation.

The contribution of such a project lies in the fact that the Xerox utilities are proprietary; compilation to FSA will enable us to use grammars developed with the Xerox tools on publicly available systems. Furthermore, parallel investigation of two similar, yet different, systems, is likely to result in new insights regarding the two systems and there interrelationships. Finally, such a compiler will enable us to compare the performance of the two systems on very similar benchmarks.

2.2 Machine-learning algorithms of Hebrew morphology

In these projects we will evaluate the applicability of several state-of-the-art machine learning algorithms to the problem of learning Hebrew morphology. Machine Learning is a general term for a variety of algorithms which improve their behavior the more times they are executed. Such algorithms can be unsupervised, which means they can only learn from the data they are executed on; or supervised, which means that they have access to other sources of knowledge.

In recent years, machine learning was extensively applied to natural language processing problems. Simple classification tasks, such as part-of-speech tagging, can be very efficiently solved using such technology. Other problems, such as word segmentation or morphological analysis, are addressed in the literature, but the performance of ML algorithms for the more complicated problems is still insufficient.
The goal of the project will be, for a particular ML algorithm, to evaluate its applicability to the problem of Hebrew morphological analysis. You will be expected to:

- read a paper describing a particular technique and understand it
- implement the algorithm described in the paper
- apply the algorithm to the data discussed in the paper
- apply the algorithm to Hebrew data
- evaluate the technique, and in particular detail the following aspects of the paper:
  - What is the problem that the technique aims to solve?
  - What resources are needed?
  - What is the technology used? Detail the proposed solution, including a full description of the algorithm.
  - What is the evaluation criterion? How can we know whether the technique "works" or not?
  - Evaluate the performance of the algorithm. Perform a complete analysis (in terms of recall and precision) of its performance.
  - Suggest ideas for improvement.

The articles we will implement are:

- Yarowsky and Wicentowski (2000)
- Schone and Jurafsky (2000)
- Goldsmith (2001)
- Creutz and Lagus (2002)
- Oflazer, Nirenburg, and McShane (2001), and extensions suggested by Zajac (2001)

3 Administration

We will meet occasionally during the semester for introduction, presentations of the papers, progress reviews and final presentations. Attendance is mandatory. All meetings will be held on Mondays. Wednesdays will be reserved for individual support and advice.

Schedule:

16.10: Introduction and assignment of projects.

6.11–13.11: Presentation of the articles. By this date you should have read and fully understood the paper, so you can present it to the class.

27.11: Progress report. Discussion of required resources, format of the input and output, problems encountered etc. By this date you should have a full design of the project.

1.1: Demonstration. By this date you should have a fully functional system, including evaluation data on English and an understanding of how to evaluate if on Hebrew.

8.1: Final submission. All documentation, including evaluation on Hebrew, must be ready by this date.

Grading will be based on comprehension of the assigned article, quality of the implementation, quality of the evaluation and innovation of the application to Hebrew. In particular, the final grade will be based on:

- Comprehension and presentation of the paper
- Full implementation of the algorithm
- Presentation of a final working system
- Comprehensive documentation
- Analysis of the results, in particular applicability for Hebrew

References

Creutz, Mathias and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *ACL'02 Workshop on Morphological and Phonological Learning*. ACL, July.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *computational Linguistics*, 27(2):153–198, June.

Oflazer, Kemal, Sergei Nirenburg, and Marjorie McShane. 2001. Bootstrapping morphological analyzers by combining human elicitation and machine learning. *computational Linguistics*, 27(1).

Schone, Patrick and Daniel Jurafsky. 2000. Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of CoNLL-2000 and LLL-2000*, pages 67–72, Lisbon, Portugal.

Yarowsky, David and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of ACL-2000*, pages 207–216, Hong Kong.

Zajac, Rémi. 2001. Morpholog: Constrained and supervised learning of morphology. In *Proceedings of CoNLL-2001*, pages 90–96, Toulouse, France.