Genomic Entity Recognition at TREC

Dina Demner-Fushman, Philip Resnik, and Douglas W. Oard
Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742
(demner,resnik,oard)@umiacs.umd.edu

1 Introduction

The University of Maryland has been as active participant in TREC, NTCIR and CLEF for many years, both as participants and as a track coordinator. In this position paper we argue for *genomic entity recognition* as a task in the TREC genomics pre-track.

Experience has shown that identification of relevant entities is a critical component in many tasks that rely on processing natural language. In the context of information extraction, for example, the message understanding (MUC) community found that named entity identification was a critical prerequisite to effective template filling. In cross-language information retrieval, named entity transliteration has been found to be important in systems that must handle different writing systems. And in machine translation, handling of named entities is under consideration as an evaluation subtask for participants in the DARPA TIDES program. Judging by our experience so far (discussed below), we believe that recognition of genomic entities is going to be an essential part of any task that involved manipulating genomic information in text. Moreover, when recognized terms are ambiguous, *disambiguation* will also be important.

We see our proposal as something that can bridge the gap between component-oriented approach and task-oriented approaches. A component orientation is dominant in the computational linguistics community, with enabling technologies like part-of-speech tagging, parsing, and word sense disambiguation typically evaluated outside the context of a broader task. A task-oriented perspective is evident in the new "KDD Cup" (http://www.biostat.wisc.edu/∼craven/kddcup/), where the tasks are defined in terms of their end-result characteristics (in this case, automating the curator's task for FlyBase), but the range of component technologies is left wide open.

2 Previous Work

Beyond our experience with large-scale evaluations in TREC-like settings, we have been involved in three relevant pieces of work: preliminary studies on automatic processing of medical research articles to help identify synonymous gene identifiers; automatic term recognition for medical text categorization based on MeSH controlled vocabulary; and development of evaluation metrics for lexical disambiguation. In this section, we briefly access each in turn.

First, during the last year, we have begun working in collaboration with Patrick Paroubek (CNRS/LIMSI) on the problem of identifying gene name synonyms in order to help connect structured databases with free-text data. Initial efforts have demonstrated the importance of proper domain-specific tokenization and other corpus-preparation steps, and they have also illustrated the rampant ambiguity present in MEDLINE articles, particularly between genes and proteins(c.f., [1]).

Second, we have been working on the terminology issues relevant to automatically categorizing MEDLINE citations according to Medical Subject Headings (MeSH), a task motivated by the rapid growth in available machine-readable unannotated publications, and by with the cost and inconsistency human indexing. In a recent study, our goal was to explore to what extent statistical similarity measures, combined

with an available controlled vocabulary thesaurus, could be used for key-term extraction for automatic and supervised indexing and as a preliminary step in text categorization. Descriptors assigned to documents by human indexers were used as a gold standard in evaluating our automatic MeSH assignments, and the UMLS metathesaurus served as the basis for evaluating term identification. We experimented with terms identified using six statistical measures of bigram association: Fisher's exact test, the log likelihood ratio, chi-squared, mutual information, Dice's coefficient, and a new affinity measure. We found that although there was no one best association measure for all cases, terms selected using a combination of our new affinity measure and log likelihood enabled us to propose a MeSH category for up to 90% of the documents, with reasonably high accuracy depending on the degree of coverage selected.

Third, we have significant experience with both supervised and automatic word sense disambiguation, and with the development of WSD evaluation measures [2, 3]. Our experience with combining corpus-based and knowledge-based approaches for lexical tasks seems seems particularly relevant in a domain where significant lexical domain knowledge has been encoded in widely available on-line knowledge sources.

3 Evaluation Proposal

The modular approach we have suggested calls for a situated evaluation in which component-level measures are used to evaluate term recognition and disambiguation in the context of a specific application. The key idea is to report both component-level measures (perhaps using MUC's F measure for genomic entity tagging accuracy, for example) and task-level measures (such as mean average precision for ranked retrieval). With both types of measures available, we expect that participants would gain additional insight into the effect of the critical components on task performance. Of course, such an approach would involves some additional work—at the very least establishing data formats and developing scoring software.

For disambiguation, a useful starting point for discussion might be the measures adopted in the SENSEVAL word sense disambiguation evaluations. These were based on proposals for evaluation of probabilistic classification into a term hierarchy [2, 3], a task abstraction that can be adapted to many settings. The basic idea is to replace the "exact match" criterion with a paradigm in which systems offer probability distributions over the set of possible classifications; the evaluation measure is defined as:

$$-\frac{1}{N}\sum_{i=1}^{N}\log_2 p_{\mathcal{A}}(c_i|w_i,\text{context}_i),$$

where N is the number of test instances and $p_{\mathcal{A}}$ is the probability assigned by algorithm $\mathcal{A}$ to the correct class, c_i, of ambiguous term w_i in context_i.

4 Conclusion

Entity recognition and disambiguation are crucial building blocks for many tasks, including classification, information retrieval, and text data mining. We believe that a situated evaluation of genomic entity recognition could serve as a useful middle ground between component-level and system-level evaluations, and we look forward to exploring how our ideas might be applied in the TREC genomic track.

References

[1] Vasileios Hatzivassiloglou, Pablo A. Duboue, and Andrey Rzhetsky. Disambiguating proteins, genes, and rna in text: A machine learning approach. In *Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology*, Tivoli Gardens, Denmark, July 2001.

[2] I. Dan Melamed and Philip Resnik. Evaluation of sense disambiguation given hierarchical tag sets. *Computers and the Humanities*, (1–2), 2000.

[3] Philip Resnik and David Yarowsky. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133, 1999.