Adaptive Information Extraction for Complex Biomedical Tasks

Donghui Feng **Gully Burns** **Eduard Hovy**
Information Sciences Insitute
University of Southern California
Marina del Rey, CA, 90292
{donghui, burns, hovy}@isi.edu

Abstract

Biomedical information extraction tasks are often more complex and contain uncertainty at each step during problem solving processes. We present an adaptive information extraction framework and demonstrate how to explore uncertainty using feedback integration.

1 Adaptive Information Extraction

Biomedical information extraction (IE) tasks are often more complex and contain uncertainty at each step during problem solving processes.

When in the first place the desired information is not easy to define and to annotate (even by humans), iterative IE cycles are to be expected. There might be gaps between the domain knowledge representation and computer processing ability. Domain knowledge might be hard to represent in a clear format easy for computers to process. Computer scientists may need time to understand the inherent characteristics of domain problems so as to find effective approaches to solve them. All these issues mandate a more expressive IE process.

In these situations, the traditional, straightforward, and one-pass problem-solving procedure, consisting of definition-learning-testing, is no longer adequate for the solution.

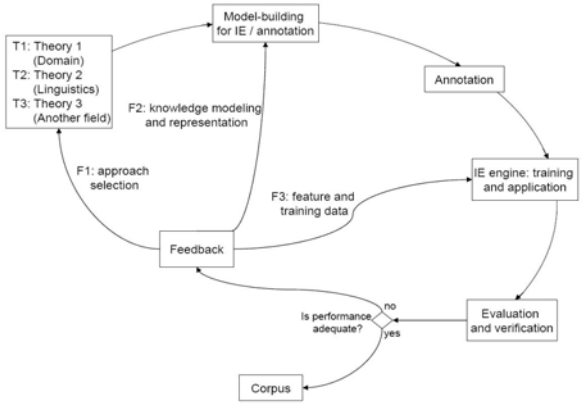

Figure 1. Adaptive information extraction.

For more complex tasks requiring iterative cycles, an adaptive and extended IE framework has not yet been fully defined although variants have been explored. We describe an adaptive IE framework to characterize the activities involved in complex IE tasks. Figure 1 depicts the adaptive information extraction framework.

This procedure emphasizes one important adaptive step between the learning and application phases. If the IE result is not adequate, some adaptations are required:

Our study focuses on extracting tract-tracing experiments (Swanson, 2004) from neuroscience articles. The goal of tract-tracing experiment is to chart the interconnectivity of the brain by injecting tracer chemicals into a region of the brain and then identifying corresponding labeled regions where the tracer is transported to (Burns *et al*., 2007). Our work is performed in the context of NeuroScholar[1], a project that aims to develop a Knowledge Base Management System to benefit neuroscience research.

We show how this new framework evolves to meet the demands of the more complex scenario of biomedical text mining.

2 Feedback Integration

This task requires finding the knowledge describing one or more experiments within an article as well as identifying desired fields within individual sentences. Significant complexity arises from the presence of a variable number of records (experiments) in a single research article --- anywhere from one to many.

Experiment	
tracerChemical	null
injectionLocation	the contralateral AVCN
labelingLocation	the DCN
labelingDescription	no labelled cells

Table 1. An example tract-tracing experiment.

Table 1 provides an example of a tract-tracing experiment. In this experiment, when the tracer was injected into the injection location “the contralateral AVCN”, “no labeled cells” was found in the labeling location “the DCN”.

For sentence level fields labeling, the performance of F1 score is around 0.79 (Feng *et al*., 2008).

[1] http://www.neuroscholar.org/

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 120–121,
Columbus, Ohio, USA, June 2008.

We here show how the adaptive information extraction framework is applied to labeling individual sentences. Please see Feng *et al*. (2007) for the details of segmenting data records.

2.1 Choosing Learning Approach via F1

A natural way to label sentences is to obtain (by hand or learning) patterns characterizing each field (Feng *et al*., 2006; Ravichandran and Hovy, 2002). We tried to annotate field values for the biomedical data, but we found few intuitive clues that rich surface text patterns could be learned with this corpus.

This insight, Feedback F1, caused us to give up the idea of learning surface text patterns as usual, and switch to the Conditional Random Fields (CRF) (Lafferty *et al*., 2001) for labeling sentences instead. In contrast to fixed-order patterns, the CRF model provides a compact way to integrate different types of features for sequential labeling problems and can reach state-of-the-art level performance.

2.2 Determining Knowledge Schema via F2

In the first place, it is not clear what granularity of knowledge/information can be extracted from text and whether the knowledge representation is suitable for computer processing. We tried a series of approaches, using different levels of granularity and description, in order to obtain formulation suitable for IE. Figure 2 represents the evolution of the knowledge schema in our repeated activities.

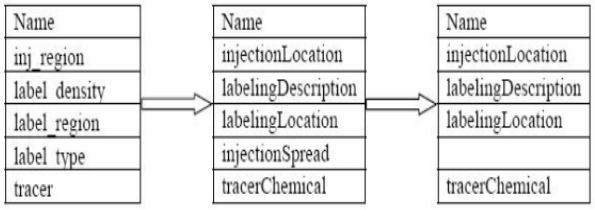

Figure 2. Knowledge schema evolution.

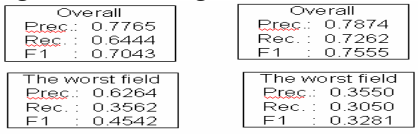

Figure 3. System performance at stage 1 and 2.

We initially started with the schema in the leftmost column but our pilot study showed that some fields, for example, "label_type", had too many variations in text description, making it very hard for CRF to learn clues about it. We then switched to the second schema but ended up seeing that the field "injectionSpread" needed more domain knowledge and was therefore not able to be learned by the systems. The last column is the final schema after those pilot studies. Figure 3 shows system performance (overall and the worst field) corresponding to the first and the second representation schemas.

2.3 Exploring Features via F3

To train CRF sentence labeling systems, it is vital to decide what features to use and how to prepare those features. Through the cycle of Feedback F3, we explored five categories of features and their combinations to determine the best features for optimal system performance. Table 2 shows system performance with different feature combinations.

System Features	Prec.	Recall	F_Score
Baseline	0.4067	0.1761	0.2458
Lexicon	0.5998	0.3734	0.4602
Lexicon + Surface Words	0.7663	0.7302	0.7478
Lexicon + Surface Words + Context Window	0.7717	0.7279	0.7491
Lexicon + Surface Words + Context Window + Window Words	0.8076	0.7451	0.7751
Lexicon + Surface Words + Context Window + Window Words + Dependency Features	**0.7991**	**0.7828**	**0.7909**

Table 2. Precision, Recall, and F_Score for labeling.

Please see Feng *et al*. (2008) for the details of the sentence level extraction and feature preparation,

3 Conclusions

In this paper, we have shown an adaptive information extraction framework for complex biomedical tasks. Using the iterative development cycle, we have been able to explore uncertainty at different levels using feedback integration.

References

Burns, G., Feng, D., and Hovy, E.H. 2007. Intelligent Approaches to Mining the Primary Research Literature: Techniques, Systems, and Examples. Book Chapter in *Computational Intelligence in Bioinformatics*, Springer-Verlag, Germany.

Feng, D., Burns, G., and Hovy, E.H. 2007. Extracting Data Records from Unstructured Biomedical Full Text. In *Proc. of EMNLP 2007*.

Feng, D., Burns, G., Zhu, J., and Hovy, E.H. 2008. Towards Automated Semantic Analysis on Biomedical Research Articles. In *Proc. of IJCNLP-2008*. Poster Paper.

Feng, D., Ravichandran, D., and Hovy, E.H. 2006. Mining and reranking for answering biographical queries on the web. In *Proc. of AAAI-2006*. pp. 1283-1288.

Lafferty, J., McCallum, A. and Pereira, F. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML-2001*.

Ravichandran, D. and Hovy, E.H. 2002. Learning surface text patterns for a question answering system. In *Proceedings of ACL-2002*.

Swanson, L.W. 2004. *Brain maps: structure of the rat brain*. 3rd edition, Elsevier Academic Press.