Notes on Identifying Relations

Ewan Klein

May 25, 2005

1 Introduction

Most (but not all) approaches to Relation Extraction (RE) assume that Named Entity Recognition has been carried out in a prior phase. This is both useful and often necessary, since most parsers would fail to correctly recognize the boundaries of biomedical terms. [Uramoto et al., 2004] do this using dictionary lookup with a two million entry dictionary, and also carry out term grounding/normalization as a pre-processing step.

2 Rule-based Patterns

[Blaschke and Valencia, 2002] uses a suite of hand-written 'frames' to recognize protein-protein interactions. Pre-processing the input text consists of labeling parts of speech and classifiying entities. Two examples are shown below.

```
[proteins] (0-5) [verbs] (0-5) [proteins]
[verbs] of (0-3) [proteins] (0-3) by (0-3) [proteins]
```

Parentheses indicate how many words can appear in that position.

Hobbs [2004] assumes that NP chunks have been identified and labeled, and that VP chunks and PPs have also been identified.

```
NG: [ FAP68 ]
VG: [ interacts specifically ]
PP: [ with
   NG : [ the inactive form of HGF receptor ]
],
such as a kinase-defective receptor or a dephosphorylated wild type
receptor.
```

It seems that patterns are defined over these shallow parse structures. An possible pattern would be

```
   <Protein> interacts with <Protein>
```

Presumably there is some kind of wild card mechanism which would actually look more like:

```
<Protein> * interacts * with <Protein>
```

A similar approach is adopted by [Thomas et al., 2000], except that the latter are more explicit about using syntactic labels in their patterns.

2.1 Pattern Transformations

One of the limitations of this approach is that there is considerable linguistic variation in how even simple relations are expressed. Hobbs addresses this issue by defining "compile-time transformations" that map the simple active form of clauses into passives, relative clauses, nominalizations, etc. Such a transformation would automatically supplement the above pattern with further patterns such as:

```
<Protein> which interacts with <Protein>
interaction of <Protein> with <Protein>
```

The patterns would be used to populate templates of a standard kind. E.g.,

```
Interaction:
ID:             Rel1
Interactor:     P1
Interactee:     P2

Protein:
ID:             P1
Name:           FAP68

Protein:
ID:             P2
Name:           HGF receptor
```

3 Probabilistic Approaches

3.1 Hierarchical HMMs

[Skounakis et al., 2003] adopt a supervised learning approach using Hierarchical Hidden Markov Models. The training data is chunked, and chunks are also labeled with NER classes. The chunk labels are determined by the chunker (Sundance) while the labels are domain specific.

```
NP_SEGMENT                      DET                 this
                                 UNK                 enzyme
NP_SEGMENT: Protein             UNK: Protein        ubc6
VP_SEGMENT:                     V                   localizes
PP_SEGMENT                      PREP                to
NP_SEGMENT: Location            ART                 the
                                 N: Location         endoplasmic
                                 N: Location         reticulum
```

The semantic relation recovered is represented as:

```
subcellular-localization(UBC6, endoplasmic reticulum)
```

The Hierarchical HMMs are like phrasal transition networks in form. Sentences are represented as sequences of phrases. The top level of the HMM emits phrases, and the lower level emits sequences of words. That is, the word level HMMs are embedded in the states of the phrase level HMMs.

3.2 Interfillers

[Bunescu et al., 2004] explore a number of surface-oriented approaches. One of them tries to learn *interfillers*. This is the text fragment that stands between two tagged entities standing in a relationship:

```
SHPTP2: interactor
[interacts with another signaling protein ,]: interfiller
Grb7: interactee
```

A related learning method, Extraction using Longest Common Subsequences (ELCS), learns rules involving disjunctions of words and bounded wild cards.

```
[These | Here | have | , ] *{,4}
[data | we | previously | the | wild] *{,1}
[suggest | show | reported | transcription | - ] *{,2}
[that | factor | type | of] *{,15}
PROT *{,14}
[surface | of | - | with | bound | activate]
PROT *{,15}
.
```

The wild cards indicate the number of unconstrained words allowed between two keywords; this approach is similar to that of [Blaschke and Valencia, 2002] except that the rules are learned, and involve words rather than POS tags.

They discuss three learning methods for ELCS, all of which start with a small number of determinate seed rules and then try to merge them into more generalized forms until precision starts to decrease.

References

C. Blaschke and A. Valencia. The frame-based module of the suiseki information extraction system. *IEEE Intelligent Systems*, (17):14–20, 2002.

R. Bunescu, R. Ge, R.J. Kate, E.M. Marcotte, R.J. Mooney, A.K. Ramani, and Y.W. Wong. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine (Special Issue on Summarization and Information Extraction from Medical Documents)*, 2004.

Jerry R. Hobbs. Information extraction from biomedical text. *Journal of Biomedical Informatics*, 2004.

M. Skounakis, M. Craven, and S. Ray. Hierarchical Hidden Markov Models for information extraction. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico. Morgan Kaufmann., 2003.*, 2003. URL `citeseer.ist.psu.edu/skounakis03hierarchical.html`.

James Thomas, David Milward, Christos Ouzounis, Stephen Pulman, and Mark Carroll. Automatic extraction of protein interactions from scientific abstracts. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 538–549, 2000.

N. Uramoto, H. Matsuzawa, T. Nagano, A. Murakami, H. Takeuchi, , and K. Takeda. A text-mining system for knowledge discovery from biomedical documents. *IBM Systems Journal*, 43 (3):516–533, 2004.