Information Extraction from Biomedical Text

Jerry R. Hobbs
Artificial Intelligence Center
SRI International
Menlo Park, California 94025
hobbs@ai.sri.com

Abstract

Information extraction is the process of scanning text for information relevant to some interest, including extracting entities, relations, and events. It requires deeper analysis than key word searches, but its aims fall short of the very hard and long-term problem of full text understanding. Information extraction represents a midpoint on this spectrum, where the aim is to capture structured information without sacrificing feasibility.

One of the key ideas in this technology is to separate processing into several stages, in cascaded finite-state transducers. The earlier stages recognize smaller linguistic objects and work in a largely domain-independent fashion. The later stages take these linguistic objects as input and find domain-dependent patterns among them.

There are now initial efforts to apply this technology to biomedical text, In other domains, the technology plateaued at about 60% recall and precision. Even if applications to biomedical text do no better than this, they could still prove to be of immense help to curatorial activities.

1 Introduction

Information extraction is the process of scanning text for information relevant to some interest, including extracting entities, relations, and, most challenging, events—or who did what to whom. It requires deeper analysis than key word searches, but its aims fall short of the very hard and long-term problem of text understanding, where we seek to capture *all* the information in a text, along with the speakers' or writer's intention. Information extraction represents a midpoint on this spectrum, where the aim is to capture structured information without sacrificing feasibility.

Information extraction technology arose in response to the need for efficient processing of texts in specialized domains. Full-sentence parsers expended a lot of effort in trying to arrive at parses of long sentences that were not relevant to the domain, or which contained much irrelevant material, thereby increasing the chances for error. Information extraction technology, by contrast, focuses in on only the relevant parts of the text and ignores the rest.

In the last ten years, the technology of information extraction has advanced significantly. It has been applied primarily to domains of economic and military interest. There are now initial efforts to apply it to biomedical text (e.g., Humphreys et al., 2000; Thomas et al., 2000), and the time is ripe for further research.

2 Cascaded Finite-State Transducers

One of the key ideas in this technology is to separate processing into several stages, in "cascaded finite-state transducers". A finite-state automaton reads one element at a time of a sequence of elements; each element transitions the automaton into a new state, based on the type of element it is, e.g., the part of speech of a word. Some states are designated as final, and a final state is reached when the sequence of elements matches a valid pattern. In a finite-state transducer, an output entity is constructed when final states are reached, e.g., a representation of the information in a phrase. In a cascaded finite-state transducer, there are different finite-state transducers at different stages. Earlier stages will package a string of elements into something the the next stage will view as a single element.

In the approach implemented in SRI International's system called FASTUS (a slightly altered acronym of Finite-State Automaton Text Understanding System)(Hobbs et al., 1997), the earlier stages recognize smaller linguistic objects and work in a largely domain-independent fashion. They use purely linguistic knowledge to recognize that portion of the syntactic structure of the sentence that linguistic methods can determine reliably, requiring relatively little modification or augmentation as the system is moved from domain to domain. The later stages take these linguistic objects as input and find domain-dependent patterns among them.

Typically there are five levels of processing:

1. Complex Words: This includes the recognition of multiwords and proper names. In biomedicine this would include names of chemical compounds.

2. Basic Phrases: Sentences are segmented into noun groups, verb groups, and particles.

3. Complex Phrases: Complex noun groups and complex verb groups are identified.

4. Domain Patterns: The sequence of phrases produced at Level 3 is scanned for patterns of interest to the application, and when they are found, semantic structures are built that encode the information about entities and events contained in the pattern.

5. Merging Structures: Semantic structures from different parts of the text are merged if they provide information about the same entity or event.

As we progress through the five levels, larger segments of text are analyzed and structured. In each of stages 2 through 4, the input to the finite-state transducer is the sequence of chunks constructed in the previous stage.

This decomposition of the natural-language problem into levels is essential to the approach. Many systems have been built to do pattern matching on strings of words. The advances in information extraction have depended crucially on dividing that process into separate levels for recognizing phrases and recognizing patterns among the phrases. Phrases can be recognized reliably with purely syntactic information, and they provide precisely the elements that are required for stating the patterns of interest.

I will illustrate the levels of processing by describing what is done on the following sentences, from a biomedical abstract.

> gamma-Glutamyl kinase, the first enzyme of the proline biosynthetic pathway, was purified to a homogeneity from an Escherichia coli strain resistant to the proline analog 3,4-dehydroproline. The enzyme had a native molecular weight of 236,000 and was apparently comprised of six identical 40,000-dalton subunits.

In this example, we will assume we are mapping the information into a complex database of pathways, reactions, and chemical compounds, such as the EcoCyc database developed by Karp and his colleagues at SRI International (Karp et al., 19??). In this database there are Reaction objects with the attributes ID, Pathway, and Enzyme, among others, and Enzyme objects with the attributes ID, Name, Molecular-Weight, Subunit-Component, and Subunit-Number.

The five phases are as follows:

1. Complex Words: This level of processing identifies multiwords such as “gamma-Glutamyl proline”, Escherichia coli”, “3,4-dehydroproline”, and “molecular weight”.

Languages in general are very productive in the construction of short, multiword fixed phrases and proper names employing specialized micrograms. This is the level at which they are recognized. The biomedical language is especially rich in this regard; this in fact may be the biggest barrier to information extraction research in biological domains. On the other hand, medical informatics has been at the forefront of human language technology in building up terminological resources, and there is much good recent work in automating the building of the lexicons and in the techniques for recognizing biomedical terms (e.g., Ananiadou et al., 2002).

2. Basic Phrases: At Level 2 the first example sentence is segmented into the following phrases:

Enzyme Name:	gamma-Glutamyl kinase
Noun Group:	the first enzyme
Preposition:	of
Noun Group:	the proline biosynthetic pathway
Verb Group:	was purified
Preposition:	to
Noun Group:	homogeneity
Preposition:	from
Noun Group:	an Escherichia coli strain
Adjective Group:	resistant
Preposition:	to
Noun Group:	the proline analog
Noun Group:	3,4-dehydroproline

Noun groups are noun phrases up through the head noun but not including the right modifiers like prepositional phrases and relative clauses. Verb groups are head verbs with their auxilliaries. Adjective phrases are predicate adjectives together with their copulas, if present.

The noun group and verb group grammars that were implemented in FASTUS were essentially those given in the grammar of Sager (1981), converted into regular expressions.

This breakdown of phrases into nominals, verbals, and particles is a linguistic universal. Whereas the precise parts of speech that occur in any language can vary widely, every language has elements that are fundamentally nominal in character, elements that are fundamentally verbal or predicative, and particles or inflectional affixes that encode relations among the other elements.

3. Complex Phrases: At Level 3, complex noun groups and verb groups that can be recognized reliably on the basis of domain-independent, syntactic information are recognized. This includes the attachment of appositives to their head noun group,

> the proline analog 3,4-dehydroproline

and the attachment of "of" prepositional phrases to their head noun groups,

> the first enzyme of the proline biosynthetic pathway.

In the course of recognizing basic and complex phrases, entities and events of domain interest are often recognized, and the structures for these are constructed. In the sample text, an Enzyme structure is constructed for gamma-Glutamyl kinase. Corresponding to the complex noun group "gamma-Glutamyl kinase, the first enzyme of the proline biosynthetic pathway," the following structure are built:

Reaction:

ID:	R1
Pathway:	proline
Enzyme:	E1

Enzyme:

ID:	E1
Name:	gamma-Glutamyl kinase
Molecular-Weight:	–
Subunit-Component:	–
Subunit-Number:	–

In many languages some adjuncts are more tightly bound to their head nouns than others. "Of" prepositional phrases are in this category, as are phrases headed by prepositions that the head noun subcategorizes for. The basic noun group together with these adjuncts constitutes the complex noun group. Complex verb groups are also motivated by considerations of linguistic universality. Many languages have quite elaborate mechanisms for constructing complex verbs. One example in English is the use of control verbs; "to conduct an experiment" means the same as "to experiment". Another example is the verb-particle constructions such as "set up".

4. Clause-Level Domain Patterns: In the sample text, the domain patterns

<Compound> have <Measure> of <values>

<Compound> comprised of<Compound>

are instantiated in the second sentence. These patterns result in the following Enzyme structures being built:

Enzyme:

ID:	E2
Name:	–
Molecular-Weight:	236,000
Subunit-Component:	–
Subunit-Number:	–

Enzyme:

ID:	E3
Name:	–
Molecular-Weight:	–
Subunit-Component:	E4
Subunit-Number:	6

Enzyme:

ID:	E4
Name:	–
Molecular-Weight:	40,000
Subunit-Component:	–
Subunit-Number:	–

This level corresponds to the basic clause level that characterizes all languages, the level at which in English Subject-Verb-Object (S-V-O) triples occur, and thus again corresponds to a linguistic universal. This is the level at which predicate-argument relations between verbal and nominal elements are expressed in their most basic form.

5. Merging Structures: The first four levels of processing all operate within the bounds of single sentences. The final level of processing operates over the whole discourse. Its task is to see that all the information collected about a single entity or relationship is combined into a unified whole. This is where the problem of coreference is dealt with in this approach.

The three criteria that are taken into account in determining whether two structures can be merged are the internal structure of the noun groups, nearness along some metric, and the consistency, or more generally, the compatibility of the two structures.

In the analysis of the sample text, we have produced four enzyme structures. Three of them are consistent with each other. Hence, they are merged, yielding

Enzyme:

ID:	E1
Name:	gamma-Glutamyl kinase
Molecular-Weight:	236,000
Subunit-Component:	E4
Subunit-Number:	6

The fourth is inconsistent because of the differing molecular weights and the subunit relation, and hence is not merged with the others.

The finite-state technology has sometimes been characterized as *ad hoc* and as *mere* pattern-matching. However, the approach of using a *cascade* of finite-state machines, where each level corresponds to a linguistic natural kind, reflects important universals about language. It was inspired by the remarkable fact that very diverse languages all show the same nominal element - verbal element - particle distinction and the basic phrase - complex phrase distinction. Organizing a system in this way leads to greater portability among domains and to the possibility of easier acquisition of new patterns.

3 Compile-Time Transformations

Natural language admits a great deal of variation. This means that patterns must be stated for not only the basic active form of clauses, but also passives, relative clauses, nominalizations, and so on. But these are for the most part predictable variations. Hence, we have implemented "compile-time transformations" that take basic Subject-Verb-Object patterns and transform them into linguistic variants. Thus, by specifying a pattern for

<Protein> inhibits <Reaction>

we automatically add patterns as well for

<Reaction> is inhibited by <Protein>

<Protein> which inhibits <Reaction>

<Protein> is inhibitor of <Reaction>

and so on.

When this was first implemented, it reduced the time required for specifying the patterns for a domain from weeks to less than a day.

4 Types of Specialized Domains

In our experience in non-biomedical domains there seem to be two types of applications. In the first, one can use what may be called a “noun-driven” approach. The type of an entity is highly predictive of its role in the event. In this case, it is not so necessary to get the Subject-Verb-Object relations correct. Looser patterns can be written. For example, if the only patterns we are looking for are

<Protein> inhibits <Reaction>

<Protein> promotes <Reaction>

Then the protein always fills the role of the effector and the reaction always fills the role of the effected.

In other domains, the roles of entities in events cannot be predicted from their type, but only from their syntactic place in sentences. These applications require what may be called a “verb-driven” approach. Tighter patterns must be written, and Subject-Verb-Object relations must be discovered. For example, in

<Protein> binds to <Protein>

we cannot tell from the fact that something is a protein which of the two roles it plays in the binding event.

The vast specialized and highly organized terminology of biomedicine suggests that perhaps a noun-driven approach would be adequate. The roles of entities may be very tightly constrained. On the other hand, as Friedman et al. (2001) have shown, there can be deeply nested relations in complex events, and it can be crucial to get the Subject-Verb-Object relations right, in which case a verb-driven approach is required.

5 The Limits of Information Extraction Technology

Information extraction is evaluated by two measures—recall and precision. Recall is a measure of completeness, precision of correctness. When you promise to tell the whole truth, you are promising 100% recall. When you promise to tell nothing but the truth, you are promising 100% precision.

In Message Understanding Conference (MUC) evaluations in the 1990s, systems doing name recognition achieved about 95% recall and precision, which is nearly human-level performance, and very much faster. In event recognition the performance plateaued at about 60% recall and precision.

There are several possible reasons for this. Our analysis of our results showed that the process of merging was implicated in a majority of our errors; we need better ways of doing event and relationship coreference. It could be that 60% is how much information texts "wear on their sleeves". Current technology can only extract what is explicit in texts . To get the rest of the information requires inference. A third possibility is that the distribution of linguistic phenomena simply has a very long tail. Handling the most common phenomena gets you to 60% relatively quickly. Getting to 100% then requires handling increasingly rare phenomena. A month's work gets you to 60%. Another year's work gets you to 65%. A fourth possibility is that errors multiply. If you can recognize an entity with 90% accuracy and to recognize a clause-level pattern requires recognizing four entities, then the accuracy should be $(.9)^4$ or about 60%.

This raises the interesting question of what utility there is in a 60% technology. Obviously you would not be happy with a bank statement that is 60% accurate. On the other hand, 60% accuracy in web search would be a distinct improvement. It is best to split this question into two parts—recall and precision.

If you have 60% recall, you are missing 40% of the mentions of relevant information. But there are half a million biomedical articles a year, and keeping up with them requires massive curatorial effort. 60% recall is an improvement if you would otherwise have access to much less. Moreover, recall is measured not on facts but on *mentions* of facts. If there are multiple mentions of some fact, we have multiple opportunities to capture it.

With 60% precision in a fully automatic system, then 40% of the information in your database will be wrong. You need a human in the loop. This is not necessarily a disaster. A person extracting sparse information from a massive corpus will have a much easier time discarding 40% of the entries than locating and entering 60%. Good tools would help in this as well. In addition, it may be that the usage of language in biomedical text is tightly enough constrained that precision will be higher than in the domains that have so far been the focus of efforts in informaiton extraction.

References

[1] Ananiadou, Sophia, Goran Nenadic, Dietrich Schuhmann, and Irena Spasic, 2002. "Term-Based Literature Mining from Biomedical Texts", in *Proceedings*, Intelligent Systems for Molecular Biology, ISMB, Text Data Mining SIG, Edmondton, Canada.

[2] Friedman, Carol, P. Kra, M. Krauthammer, H. Yu, A. Rzhetsky, 2001. "GENIES: A Natural-Langauge Processing System for the Extraction of Molecular Pathways from Journal Articles", *Bioinformatics* 2001:suppl1:S74-82.

[3] Hobbs, Jerry R., Douglas E. Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson, 1997. "FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text", in E. Roche and Y. Schabes, eds., *Finite State Devices*

for Natural Language Processing, MIT Press, Cambridge, Massachusetts, pp. 383-406.

[4] Humphreys, K, Demetrion G, Gaizauskas R. Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures: 2000:505-516.

[5] Karp, Peter D., 2001. "Pathway Databases: A Case Study in Computational Symbolic Theories", *Science*, vol. 293, pp. 2040-2044.

[6] Sager, Naomi, 1981. *Natural Language Information Processing: A Computer Grammar of English and Its Applications*, Addison-Wesley, Reading, Massachusetts.

[7] Thomas, J, Milward, D, Ouzounis C, Pulman S, Carroll M. Automatic extraction of protein interactions from scientific abstracts. Pac Symp Biocomput 2000; 517-528.