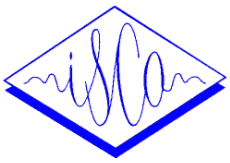

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13–16, 2003

DISCOVERY METHODS FOR INFORMATION EXTRACTION

Ralph Grishman

New York University
New York, NY
grishman@cs.nyu.edu

1. INTRODUCTION

Information extraction (IE) involves automatically identifying instances of a specified type of relation or event in text, and collecting the arguments and modifiers of the relation/event. High quality, easily adaptable IE systems would have a major effect on the ways in which we can make use of information in text (and ultimately, in speech as well).

At the present state of the art, however, performance varies widely depending on the nature of the language being processed and the complexity of the relation being extracted. For restricted sublanguages and simple relations, levels of accuracy comparable to human coders are possible. This has been achieved, for example, for some types of medical records, where both physicians and an extraction system identified diseases with 70-80% accuracy (Friedman et al. 1995). High performance has also been achieved for semi-structured Web documents — documents with some explicit mark-up (Cohen and Jensen 2001). In contrast, for more complex relations and more general texts, accuracies of 50-60% are more typical. Even at these levels IE can be of significant value in situations where the text is too voluminous to be reviewed manually; for example, to provide a document search tool much richer than current keyword systems (Grishman et al 2002). IE is also being used in other applications where perfect recall is not required, such as data mining from text collections and the generation of time lines for texts.

To make IE a more widely-useable technology, we face a two-fold challenge: improving its performance and improving its portability to new domains. Our group, and other research groups, are exploring how corpus-based training methods can address these challenges. The difficulty of IE lies in part in the wide variety of ways in which a given relation may be expressed. Automated tools for corpus analysis can help in analyzing large corpora to find these varied expressions, and hopefully can find a wider range of expressions with less human effort than current methods.

2. STRUCTURE OF AN IE SYSTEM

To understand the challenge more clearly, we need to examine the structure of current IE systems. In simplest terms, there are three stages of processing:

- linguistic pre-analysis
- IE pattern matching
- anaphora and predicate merging

The input text is first subject to a certain degree of general linguistic analysis. The amount of analysis varies among systems and among languages; a deeper analysis simplifies the next stage of processing, but may introduce errors from which it is difficult to recover. Most systems do dictionary look-up and/or part-of-speech tagging; name identification; and identification of at least simple noun groups. Some systems perform more extensive syntactic analysis, analyzing clauses or possibly the entire sentence.

The stage of 'IE pattern matching' looks for linguistic structures indicative of the relation or event to be extracted. The form of these patterns depends on the degree of linguistic pre-analysis. If only lexical and name analysis has been performed, this will typically be a regular expression involving lexical items. If partial syntactic analysis was done, the regular expressions will involve these constituents. If a full syntactic analysis was made, the patterns will involve relations of syntactic structure rather than contiguity.

In some simple applications, one document will report exactly one event or relation, so there is no issue of associating an argument with the proper event. More typically, however, a document may describe several events, and the information about a single event may be scattered across the document. In such cases the successful merging of this scattered information may be as crucial as the initial identification of the partial information.

3. PATTERN DISCOVERY

The success of IE pattern matching depends heavily on the completeness of the pattern set being used. As in many other linguistic tasks, collecting relatively complete sets is difficult because of the long tail of the distribution: there will be a few common patterns and a

large number of less frequent ones. This problem has led to the study of discovery methods for these patterns.

There has been considerable work on the supervised learning of extraction patterns, using corpora which have been annotated to indicate the information to be extracted (Califf and Mooney 1999; Soderland 1999). A range of extraction models have been used, including both symbolic rules and statistical rules such as HMMs. These methods have been particularly successful at analyzing semi-structured text, in which short passages of text appear with explicit labels or mark-up, as is often the case on the Web. In these cases relatively simple patterns can yield high extraction performance. For more complex relations and less restricted text, however, the variation in linguistic form is greater. Accordingly, the patterns or rules are more complex and much more annotated data is needed to train the model.[1] However, marking large amounts of text data for complex relationships is very time consuming (and expensive). This may make it difficult to push performance significantly using supervised methods.

The limitations of supervised methods have led to the consideration of (nearly) unsupervised methods for finding patterns. What evidence can we use to find patterns? The most promising approach to date has been the distribution of patterns in relevant vs. irrelevant documents. Patterns which occur relatively more frequently in documents which are relevant to the extraction task than in other documents are very likely to be significant patterns.

This heuristic was initially exploited by Riloff (1996) on the MUC-3/4 terrorist corpus, which has over 1300 documents hand-classified into relevant and irrelevant sets. Even marking relevance judgements, however, can be a significant effort for a large corpus. Yangarber (Yangarber et al. 2000) extended this to a bootstrapping approach which acquired both a corpus of relevant documents and a set of patterns in tandem. The training corpus is pre-processed with a named-entity tagger and a parser to identify all the subject-verb-object patterns in the corpus. Starting with a small set of 'seed' patterns that are known to be relevant to the task, the procedure retrieves documents containing these patterns. It then ranks the patterns with respect to their relative frequency in the retrieved documents and the remaining documents. The top-ranked patterns are added to the seed set and the process is repeated. Yangarber demonstrated an effectiveness at finding relevant patterns comparable to that achieved through manual text analysis. Sudo (Sudo et al. 2001) took a somewhat different approach in a system for Japanese extraction; the starting point is a *topic description* for the extraction task. A set of relevant documents is retrieved using an information retrieval system, and then patterns are ranked based on relative frequencies in the retrieved documents and the entire corpus. We may hope that future refinements of these methods, which in principle can mine very large document collections, will allow us to outperform current manually-prepared pattern sets.

[1] For example, the Univ. of Massachusetts system, which employed automated pattern collection with manual review, obtained performance comparable to manual pattern development on the 1300-document MUC-4 corpus, but significantly lower performance on the 100-document MUC-6 corpus (Fisher et al. 1995).

3. 1. Paraphrase Discovery for Patterns

These document-relevance based methods have both the strength and the shortcoming of grouping together all the predications related to a topic. This can be a benefit (compared, say, to methods which acquire all the forms of expression of a specified relation) if we want to gather all the important facts about a topic, but do not recognize that a particular relation is important for this topic. On the other hand, it means that an additional step is required to sort out acquired patterns which express very different relationships (for example, in the executive succession domain, to distinguish hiring from firing; in the medical domain, to distinguish patients who recover from those who die).

To address this problem, other researchers are investigating methods which specifically acquire paraphrase relations. These approaches start with two or more texts which report the same information. They then attempt to align passages within the text which involve the same individuals or objects and propose these as paraphrases.

Barzilay and McKeown (2001) applied this approach to multiple translations of the same foreign-language book (though not for information extraction purposes). Such sources yield closely parallel texts. Shinyama et al. (2002) used multiple articles on the same news event (identified automatically using Topic Detection and Tracking methods); these vary much more widely in structure. Given two articles, a second alignment phase identified pairs of sentences involving the same named entities (named people, organizations, locations, dates). Finally, given parsed pairs of sentences, parse tree alignment attempted to identify potential paraphrase structures. This procedure was applied to articles on two topics – management succession and crime reports. To reduce noise (incorrect paraphrase identification), consideration was limited to potential extraction patterns for these domains – to structures which appeared relatively more frequently in the articles on this topic. Moderate success was reported in the management succession domain, but much larger-scale experiments are required.

4. WORD CLASS DISCOVERY

Word classes are tightly intertwined with patterns, both for extraction and for pattern discovery.

In simple cases, patterns can be stated in terms of specific lexical items. Most often, however, that will not be sufficient; patterns must involve word classes as well. For example, if we are collecting instances of

murders, we might use the pattern 'X shot Y', which would be OK for 'Fred shot Harry' but not for 'Fred shot a roll of film'; thus a more specific pattern such as 'shot *person*' is required. Stating patterns in terms of word classes means, in turn, that the performance of pattern matching (and hence of the whole IE system) depends on the system's ability to identify instances of the word class.

Furthermore, it is difficult to perform pattern discovery without at least some word classes. If there is no notion of word classes, subject-verb-object structures must be stated in terms of specific words, so there will be little repetition and no meaningful frequency statistics for units larger than individual words. Most of the cited systems used at least a named entity tagger, which made it possible to generalize from particular names to the classes *person-name*, *organization-name*, etc.

Thus both successful discovery and successful extraction depend on relatively complete sets of word classes. There is a long history of research on the acquisition of word classes, and there has been renewed interest in connection with the needs of IE. Most of this work has involved unsupervised learning.[2] The basic idea of word class discovery is that words in similar contexts are similar, and should be placed in the same word class. Thus a typical procedure will begin with a small set of terms (a 'seed') known to be in a category. It will look for contexts which occur frequently with such words, and then find other words appearing in the same contexts, gradually building up a cluster around the seed. Within this general approach, there are a broad range of procedures, differing in several regards:

1. **the type of items classified**: Some researchers have looked specifically at classifying names (Collins and Singer 1999; Cucerzan and Yarowsky 1999). Name classification is appealing because most names (in general texts) fall into a small number of categories (such as people, places, organizations, products) and there is relatively little ambiguity for full names (although the abbreviated names which appear subsequently in a text may be ambiguous). Furthermore, names typically can be classified based both on internal evidence (e.g., begins with "Fred" or ends with "Associates") and external evidence (is followed by "died"). A great deal of work has been done on classification of common nouns, appearing as the heads of noun phrases (Riloff and Jones 1999; Thelen and Riloff 2002; Roark and Charniak 1998). Nouns can be easily identified syntactically, but in the general language they fall into a wide range of classes, with less sharp class divisions and more frequent homographs. For some technical domains it is necessary to identify and classify multi-word terms as well as single words (Yangarber et al. 2002); the identification of these terms has been a goal of terminology research.
2. **the types of contexts considered**: Some methods use as contexts the immediately adjacent words, within a window of 1 to 3 words. Other methods take account of syntactic structure, and use the governing predicate, the arguments and modifiers, and coordinated elements. In principle the words in syntactic relations are better indicators, but errors in syntactic analysis can contribute noise. Some approaches use only selected syntactic relations which can be acquired more accurately or easily (e.g., by finite-state rules).
3. **how the similarity is computed**: Given words which appear in several shared contexts, there are many variations possible in computing the similarity. In particular, most methods assign scores to the contexts: what fraction of the terms in a given context are known to be of the category of interest; (if we have negative evidence, or are acquiring multiple categories) what fraction are known to be of a different category. Given these scores, the patterns may be ranked and the procedure may use just the best context, or the best N contexts, to find additional category members. Alternatively, all contexts may be used, with weights depending on these scores, and a rule for combining evidence from multiple contexts. Note that in an iterative procedure, the scores for the contexts will be recomputed once items have been added to the cluster.

Further variations are possible in how the clusters are built. One can build one cluster at a time, or multiple 'competing' clusters. The latter has the benefit of bounding the growth of individual clusters at the borders with other clusters, thus improving precision at high recall levels (Yangarber et al. 2002; Thelen and Riloff 2002). Cluster membership can be binary (in/out), or can be graded, with the degree of membership based on the strength of the evidence. Using graded membership during acquisition may yield better clusters, even if the final result is reported in binary terms.

Although these methods have generally been presented as unsupervised learners, they can potentially also be used as active learners, where a user is part of the loop, reviewing each proposed member of the word class before it is added to the cluster.

5. AGGREGATION AND ANAPHORA

The extraction patterns, using the word classes, should be able to identify instances of the relations or events of interest appearing in the text. This, however, is not sufficient to properly identify the events with their

[2] The main exception being the work on named entity taggers. The goal here is to be able to classify new names as well as previously seen ones, and most of the systems have been trained on annotated corpora.

arguments, because information about a single event may be scattered among several sentences of the document. The system needs to be able to handle:

- explicit anaphoric elements: Fred Smith was invited to dinner. *He* was going to be promoted to president.
- arguments and modifiers which must be recovered from a larger context because they appear outside the scope of the extraction pattern: Several promotions were announced *last year*. Fred Smith was named president, …
- multiple descriptions of an event, each of which may provide partial information

Properly addressing these various *discourse* phenomena has proven to be difficult, but it will be critical to the development of high-performance IE.

Pronominal anaphora has been extensively studied (both in linguistics and computational linguistics); common noun phrase anaphors less so. Relatively simple algorithms, based on position, number, and gender, do moderately well for pronouns; baseline performance for common noun phrase anaphors, based on determiner and head, is less satisfactory. There have been a number of experiments using machine learning on coreference-annotated corpora to improve performance, but relatively little gain has been achieved so far for general language texts.[3] As annotated corpora get larger, we can expect some further improvements, but coreference-annotated corpora are expensive. A decomposition of the problems which allows components to be learned separately may be essential for significant further progress. It is not clear how much further we can go with these relatively shallow, 'knowledge-poor' methods.

Implicit arguments have generally been dealt with in two ways. Sometimes they are treated like pronominal anaphora. Other systems treat it as a special case of a more general merging problem … when should two separate pieces of information be considered part of the same event. There have been a few attempts to learn merging rules from annotated corpora (Fisher et al. 1995 (the "WRAP-UP" component); Kehler 1998; Chieu and Ng 2002), but a more systematic effort to decompose and analyze this merging task is required.

6. CONCLUSION

The need to retrieve and mine data from ever larger text collections is pushing IE into more and more applications. Nonetheless, significant problems remain both in the labor required to port IE systems to new domains and in the performance of IE systems for more complex relations. These problems are a reflection of the multiple tasks which must be successfully accomplished for IE: identifying operands of the appropriate word classes, identifying patterns corresponding to relations and events, and combining multiple pieces of information. Corpus-based methods, and in particular unsupervised learners which can take advantage of very large text collections, hold the promise of improving upon current manual analysis methods, and thus improving both the portability and performance of IE systems.

[3] See *Computational Linguistics*, Volume 27, No. 4, Special Issue on Computational Anaphora Resolution, December 2001, and (Soon et al. 2001) therein.

7. ACKNOWLEDGEMENTS

This research was supported by the Defense Advanced Research Projects Agency as part of the Translingual Information Detection, Extraction and Summarization (TIDES) program, under Grant N66001-001-1-8917 from the Space and Naval Warfare Systems Center San Diego, and by the National Science Foundation under Grant IIS-0081962. This paper does not necessarily reflect the position or the policy of the U.S. Government.

8. REFERENCES

[Barzilay and McKeown 2001] R. Barzilay and K. R. McKeown. Extracting paraphrases from a parallel corpus. *Proc. ACL/EACL 2001*.

[Califf and Mooney 1999] Mary Elaine Califf and Raymond Mooney. Relational learning of pattern-match rules for information extraction. *Proc. 16th National Conference on Artificial Intelligence (AAAI-99),* 328-334.

[Chieu and Ng 2002] Hai Leong Chieu and Hwee Tou Ng. A maximum entropy approach to IE from semi-structured and free text. *Proc. 19th National Conf. On Artificial Intelligence (AAAI-02)*.

[Cohen and Jensen 2001] William Cohen and Lee Jensen. A structured wrapper induction system for extracting information from semi-structured documents. *Workshop on Adaptive Text Extraction and Mining, 17th Int'l Joint Conf. on Artificial Intelligence,* Seattle, Wash., August, 2001.

[Collins and Singer 1999] M. Collins and Y. Singer. Unsupervised models for named entity classification. *Proc. Joint SIGDAT Conf. on EMNLP/VLC*, 1999.

[Cucerzan and Yarowsky 1999] S. Cucerzan and D. Yarowsky. Language-independent named entity recognition combining morphological and contextual evidence. *Proc. Joint SIGDAT Conf. on EMNLP/VLC*, 1999.

[Fisher at al. 1995] David Fisher, Stephen Soderland, Joseph McCarthy, Fangfang Feng, and Wendy Lehnert. Description of the UMass system as used for MUC-6. *Proc. Sixth Message Understanding Conf. (MUC-6)*. Columbia, MD, Nov. 1995.

[Friedman et al. 1995] C. Friedman, G. Hripcsak, W. DuMouchel, S. B. Johnson , and P. D. Clayton. Natural language processing in an operational clinical

information system. *Natural Language Engineering* 1995; 1:1-28.

[Grishman et al 2002] Ralph Grishman, Silja Huttunen, and Roman Yangarber. Real-time event extraction for infectious disease outbreaks. *Proc. HLT 2002 (Human Language Technology Conference)*, San Diego, California, March 2002.

[Kehler 1998] Andrew Kehler. Learning embedded discourse mechanisms for information extraction. *Proc. AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*. Stanford, CA, March 1998.

[Riloff 1996] Ellen Riloff. Automatically generating extraction patterns from untagged text. *Proc. 13th National Conf. On Artificial Intelligence (AAAI-96)*, 1044-1049.

[Riloff and Jones 1999] Ellen Riloff and Rosie Jones. Learning dictionaries for information extraction by multi-level bootstrapping. *Proc. 16th National Conf. On Artificial Intelligence (AAAI-99)*.

[Roark and Charniak 1998] Brian Roark and Eugene Charniak. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. *Proc. 36th Annl. Meeting Assn. for Computational Linguistics and 17th Int'l Conf. on Computational Linguistics (COLING-ACL '98)*, Montreal, Canada, August, 1998, 1110-1116.

[Shinyama et al 2002] Yusuke Shinyama, Satoshi Sekine, Kiyoshi Sudo, and Ralph Grishman. Automatic paraphrase acquisition from news articles. *Proc. HLT 2002 (Human Language Technology Conference)*, San Diego, California, March 2002.

[Soderland 1999] Stephen Soderland. Learning information extraction rules for semi-structured and free text. *Machine Learning*, **34**: 233-272, 1999.

[Sudo et al. 2001] Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman. Automatic pattern acquisition for Japanese information extraction. *Proc. HLT 2001 (Human Language Technology Conference)*, San Diego, CA, 2001.

[Soon et al. 2001] Wee Meng Soon, Daniel Chung Yong Lim, and Hwee Tou Ng. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics* **27**: 4, 521-544.

[Thelen and Riloff 2002] Michael Thelen and Ellen Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. *Proc. 2002 Conf. on Empirical Methods in Natural Language Processing*, Philadelphia, PA, July 2002, 214-221.

[Yangarber et al. 2000] Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. Automatic acquisition of domain knowledge for information extraction. *Proc. 18th Int'l Conf. on Computational Linguistics (COLING 2000)*, Saarbrücken, Germany, July-August 2000, 940-946.

[Yangarber 2002] Roman Yangarber, Winston Lin, and Ralph Grishman. Unsupervised learning of generalized names. *Proc. Nineteenth Int'l Conf. on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, August 2002.