Structural Linguistics and Unsupervised Information Extraction

Ralph Grishman
Dept. of Computer Science
New York University
715 Broadway, 7th floor
New York, NY 10003, USA
grishman@cs.nyu.edu

Abstract

A precondition for extracting information from large text corpora is discovering the information structures underlying the text. Progress in this direction is being made in the form of unsupervised information extraction (IE). We describe recent work in unsupervised relation extraction and compare its goals to those of grammar discovery for science sublanguages. We consider what this work on grammar discovery suggests for future directions in unsupervised IE.

1 Introduction

Vast amounts of information are available in unstructured text form. To make use of this information we need to identify the underlying information structures – the classes of entities and the predicates connecting these entities – and then automatically map the text into these information structures.

Over the past decade there has been a quickening pace of research aimed at automating the process of discovering these structures, both for broad domains (news stories) and for more limited technical domains. This has taken the form of unsupervised methods for entity set creation and relation extraction.

This research has sometimes been presented as an entirely new exploration into discovery procedures for information structures. But in fact there are relevant precedents for such exploration, albeit completely manual, going back at least half a century. An examination of these precedents can give us some insight into the challenges that face us in unsupervised information extraction.

This brief paper will first review some of the relevant work in unsupervised IE; then turn to some of the earlier work on discovery procedures within linguistics; and finally consider what these discovery procedures may suggest for the next steps in automation.

2 Unsupervised IE

The task of unsupervised information extraction involves characterizing the types of arguments that can occur with particular linguistic predicates, and identifying the predicate-argument combinations which convey the same meaning.

Most of the recent work on unsupervised IE has focused on the more tractable problem of unsupervised binary relation extraction (URE). A potential binary relation instance (a 'triple') consists of two *arguments* (typically names or nouns) connected by a *relation phrase*. A text corpus will yield a large number of such triples. The challenge is to group the triples (or a subset of them) into a set of semantically coherent relations – clusters of triples representing approximately the same semantic relationship.

One of the first such efforts was (Hasegawa et al. 2003), which used a corpus of news stories and identified the most common relations between people and companies. They relied on standard predefined named entity tags and a clustering strategy based on a simple similarity metric between

Proc. of the Joint Workshop on Automatic Knowledge Base Construction & Web-scale Knowledge Extraction (AKBC-WEKEX), pages 57–61,
NAACL-HLT, Montréal, Canada, June 7-8, 2012.

relation instances (treating relation phrases as bags of words). Subsequent research has expanded this research along several dimensions.

(Zhang et al. 2005) introduced a more elaborate similarity metric based on similarity of relation phrase constituent structure. However, this direction was not generally pursued; (Yao et al. 2011) used the dependency path between arguments as a feature in their generative model, but most URE systems have treated relation phrases as unanalyzed entities or bags of words.

Several researchers have extended the scope from news reports to the Web, using in particular robust triples extractors such as TextRunner (Banko et al. 2007) and ReVerb (Fader et al. 2011), which capture relation phrases involving verbal predicates. Moving to a nearly unrestricted domain has meant that predefined argument classes would no longer suffice, leading to systems which constructed both entity classes and relation classes. SNE (Kok and Domingos 2008) used entity similarities computed from the set of triples to construct such entity classes concurrently with the construction of relation sets. (Yao et al. 2011) learned fine-grained argument classes based on a predefined set of coarse-grained argument types. Resolver (Yates and Etzioni 2007) identified coreferential names while building relation classes.

Moving to Web-scale discovery also magnified the problem of relation-phrase polysemy, where a phrase has different senses which should be clustered with different sets of triples. This was addressed in WEBRE (Min et al. 2012), which clusters generalized triples (consisting of a relation phrase and two argument classes). WEBRE also uses a larger range of semantic resources to form the argument classes, including hyponymy relations, coordination patterns, and HTML structures.

Most of this work has been applied in the news domain or on Web documents, but there has also been some work on technical domains such as medical records (Rink and Harabagiu 2011).

Closely related to the work on unsupervised IE is the work on collecting paraphrase and inference rules from corpora (e.g., DIRT (Lin and Pantel 2001)) and some of the work on unsupervised semantic parsing (USP), which maps text or syntactically-analyzed text into logical forms. The predicate clusters of USP (Poon and Domingos 2009) capture both syntactic and semantic paraphrases of predicates. However, the work on paraphrase and USP is generally not aimed at creating an inventory of argument classes associated with particular predicates.

All of this work has been done on binary relations.[1] Discovering more general (n-ary) structures is more difficult because in most cases arguments will be optional, complicating the alignment process. (Shinyama and Sekine 2006), who built a system for unsupervised event template construction, addressed this problem in part through two levels of clustering – one level capturing stories about the same event, the second level capturing stories about the same *type* of event. (Chambers and Jurafsky 2011) clustered events to create templates for terrorist events which involve up to four distinct roles.

3 Discovery of Linguistic Structure

Current researchers on unsupervised IE share some goals with the structural linguists of the early and mid 20th century, such as Bloomfield, Saussure, and especially Zellig Harris. "The goal of structural linguistics was to discover a grammar by performing a set of operations on a corpus of data" (Newmeyer 1980, p. 6). Harris in particular pushed this effort in two directions: to discover the relations between sentences and capture them in *transformations*; and to study the grammars of *science sublanguages* (Harris 1968).

A sublanguage is the restricted form of a natural language used within a particular domain, such as medicine or a field of science or engineering. Just as the language as a whole is characterized by word classes (noun, transitive verb) and patterns of combination (noun + transitive verb + noun), the sublanguage is characterized by domain-specific word classes (N_{ion}, $V_{transport}$, N_{cell}) and patterns (N_{ion} + $V_{transport}$ + N_{cell}). Just as a speaker of the language will reject a sequence not matching one of the patterns (Cats eat fish. *but not* Cats fish eat.) so will a speaker of the sublanguage reject a sequence not matching a sublanguage pattern (Potassium enters the cell. *but not* The cell enters potassium.) Intuitively these classes and patterns have semantic content, but Harris asserted they could be characterized on a purely structural basis.

[1] Unsupervised semantic parsing (Poon and Domingos 2009) could in principle handle n-ary relations, although all the examples presented involved binary predicates. USP is also able to handle limited nested structures.

Harris described procedures for discovering the sublanguage grammar from a text corpus. In simplest terms, the word classes would be identified based on shared syntactic contexts, following the distributional hypothesis (Harris 1985; Hirschman et al. 1975). This approach is now a standard one for creating entity classes from a corpus. This will reduce the corpus to a set of word class sequences. The sublanguage grammar can then be created by aligning these sequences, mediated by the transformations of the general language (in other words, one may need to decompose and reorder the sequences using the transformations in order to get the sequences to align).

Fundamental to the sublanguage grammar are the set of elementary or *kernel* sentences, such as N_{ion} $V_{transport}$ N_{cell} ("Potassium enters the cell."). More complex sentences are built by applying a set of *operators*, some of which alter an individual sentence (e.g., passive), some of which combine sentences (e.g., conjunction, substitution). The derivation of a sentence is represented as a tree structure in which the leaves are kernel sentences and the internal nodes are operators. Within the sublanguage, repeating patterns (subtrees) of kernel sentences and operators can be identified. One can distill from these repeating derivational patterns a set of *information formats*, representing the informational constituents being combined (Sager et al. 1987).

Harris applied these methods, by hand but in great detail, to the language of immunology (Harris et al. 1989). Sager and her associates applied them to clinical (medical) narratives as well as some articles from the biomedical literature (Sager et al. 1987).

These sublanguage grammatical patterns really reflect the information structures of the language in this domain – the structures into which we want to transform the text in order to capture its information content. In that sense, they represent the ultimate goal of unsupervised IE.

Typically, they will be much richer structures than are created by current unsupervised relation extraction. For papers on lipid metabolism, for example, 7 levels of structure were identified, including kernel sentences, quantifying operators (rates, concentrations, etc.), causal relations, modal operators, and *metalanguage* relations connecting experimenters to reported results (Sager et al. 1987, p. 226).

Harris focused on narrow sublanguages because the 'semantic' constraints are more sharply drawn in the sublanguage than are seletional constraints in the language as a whole. (They were presumably also more amenable to a comprehensive manual analysis.) If we look at a broader domain, we can still expect multiple levels of information, but not a single overall structure as was possible in analyzing an individual topic.

4 What we can learn

What perspective can Harris's work provide about the future of discovery methods for IE?

First, it can give us some picture of the richer structures we will need to discover in order to adequately represent the information in the text. Current URE systems are limited to capturing a set of relations between (selected classes of) names or nouns. In terms of the hierarchical information structures produced by sublanguage analysis, they are currently focused on the predicates whose arguments are the leaves of the tree (roughly speaking, the kernel sentences).[2] For example, for

> The Government reported an increase in China's export of coal.

we would expect a URE to capture

> *export*(China, coal)

which might be able to generalize this to

> *export*(*country*, *commodity*)

but not capture the *report(Government, increase)* or *increase(export)* relations. A more comprehensive approach would also capture these and other relations that arise from modifiers on entities, including quantity and measure phrases and locatives; modifiers on predicates, including negation, aspect, quantity, and temporal information; and higher-order predicates, including sequence and causal relations and verbs of belief and reporting. The modifiers would in many cases yield unary relations; the higher-order predicates would take arguments which are themselves relations and would be generalized to relation sets.[3]

[2] This is not quite true of OpenIE triples extractors, which may incorporate multiple predicates into the relation, such as 'expect a loss' or 'plan to offer'.

[3] Decomposing predicates in this way – treating "report an increase in export" as three linked predicates rather than one – has both advantages and disadvantages in capturing paraphrase – i.e., in grouping equivalent relations. Advantages because paraphrases at a single level may be learned more easily: one doesn't have to learn the equivalence of "began to

A second benefit of studying these precedents from linguistics is that they can suggest some of the steps we may need to enrich our unsupervised IE processing. Current URE systems suffer from cluster recall problems … they fail to group together many triples which should be considered as instances of the same relation. For example, SNE (Kok and Domingos 2008) cites a pairwise recall for relations of 19%, reporting this as a big improvement over earlier systems. In Harris's terms, this corresponds to not being able to align word sequences. This is a reflection in part of the fact that most systems rely primarily or entirely on distributional information to group triples, and this is not sufficient for infrequent relation phrases.[4] Expanding to Web-scale discovery will increase the frequency of particular phrases, thus providing more evidence for clustering, but will also introduce new, complex phrases; the combined 'tail' of infrequent phrases will remain substantial. The relation phrases themselves are for many systems treated as unanalyzed word sequences. In some cases, inflected forms are reduced to base forms or phrases are treated as bags of words to improve matching. A few systems perform dependency analysis and represent the relation phrases as paths in the dependency tree. Even in such cases active objects and passive subjects are typically labeled differently, so the system must learn to align active and passive forms separately for each predicate. (In effect, the alignment is being learned at the wrong level of generality.)

The problem will be more acute when we move to learning hierarchical structures because we will have both verbal and nominalized forms of predicates and would like to identify their equivalence.

Not surprisingly, the alignment process used by Harris is based on a much richer linguistic analysis involving transformationally decomposed sentences. This included regularization of passives, normalization of nominalizations (equating "uptake of potassium by the cell" with "the cell takes up potassium"), treatment of support verbs (reducing "take a walk" to "walk"), and handling of transparent nouns (reducing "I ate a pound of chocolate" to "I ate chocolate." plus a quantity modifier on "chocolate"). Many instances of equivalent sublanguage patterns could be recognized based on such syntactic transformations.

Incorporating such transformations into an NLP pipeline produces of course a slower and more complex analysis process than used by current URE systems, but it will be essential in the long term to get adequate cluster recall – in other words, to unify different relation phrases representing the same semantic relation. Its importance will increase when we move from binary to n-ary structures. Fortunately there has been steady progress in this area over the last two decades, based on increasingly rich representations: starting with the function tags and indexing of Penn TreeBank II, which permitted regularization of passives, relatives, and infinitival subjects; PropBank, which enabled additional regularization of verbal complements (Kingsbury and Palmer 2002), and NomBank (Meyers et al. 2004) and NomLex, which support regularization between verbal and nominal constructs. These regularizations have been captured in systems such as GLARF (Meyers et al. 2009), and fostered by recent CoNLL evaluations (Hajic et al. 2009).

In addition to these transformations which can be implicitly realized through predicate-argument analysis, Harris (1989, pp. 21-23) described several other classes of transformations essential to alignment. One of these is modifier movement: modifiers which may attach to entities or arguments ("In San Diego, the weather is sunny." vs. "The weather in San Diego is sunny."). If we had a two-level representation, with both entities and predicates, this would have to be accommodated through the alignment procedure.

There will be serious obstacles to automating the full discovery process. The manual 'mechanical' process is no doubt not as mechanical as Harris would have us believe. Knowledge of the meaning of individual words surely played at least an implicit role in decisions regarding the granularity of word classes, for example. Stability-based clustering (Chen et al. 2005) has been applied to select a suitable granularity for relation clusters but the optimization will be more complex when clustering both entities and relations. Obtaining accurate syntactic structure across multiple domains will

X" and "started Xing" for each verb X. Disadvantages because paraphrases may require aligning a single predicate with two predicates, as in "grow" and "increase in size". Handling both cases may require maintaining both composite and decomposed forms of a relation.

[4] Some systems, such as WEBRE (Min et al. 2012), do use additional semantic resources and are able to achieve better recall.

require adaptive methods (for modifier attachment, for example). However, it should be possible to apply these enhancements – in structural complexity and linguistic matching – *incrementally* starting from current URE systems, thus gradually producing more powerful discovery procedures.

References

Michele Banko, Michael Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. *Proc. 20th Int'l Joint Conf. on Artificial Intelligence.*

Nathanael Chambers and Dan Jurafsky. 2011. Template-Based Information Extraction without the Templates. *Proceedings of ACL 2011.*

Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. 2005. Automatic relation extraction with model order selection and discriminative label identification. *Proc. IJCNLP 2005*.

Jan Hajic, Massimiliano Ciarmita, Richard Johansson, Daisuke Kawahara, Maria Antonia Marti, Lluis Marquez, Adam Meyers, Joakim Nivre, Sebastian Paso, Jan Stepanek, Pavel Stranak, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 Shared Task: syntactic and semantic dependencies in multiple languages. *Proc. Thirteenth Conf. on Computational Natural Language Learning (CoNLL)*, Boulder, Colorado.

Zellig Harris. 1968. *Mathematical Structures of Language.* New York: Interscience.

Zellig Harris. 1985. Distributional Structure. *The Philosophy of Linguistics*. New York: Oxford University Press.

Zellig Harris, Michael Gottfried, Thomas Ryckman, Paul Mattcik Jr., Anne Daladier, T.N. Harris and S. Harris. 1989. *The Form of Information in Science: Analysis of an Immunology Sublanguage.* Dordrecht: Kluwer Academic Publishers.

Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. 2004. Discovering Relations among Named Entities from Large Corpora. *Proceedings of ACL 2004*.

Lynette Hirschman, Ralph Grishman, and Naomi Sager. Grammatically-based automatic word class formation. 1975. *Information Processing and Management* **11**, 39-57.

P. Kingsbury and Martha Palmer. 2002. From treebank to propbank. *Proc. LREC-2002.*

Stanley Kok and Pedro Domingos. 2008. Extracting Semantic Networks from Text via Relational Clustering. *Proceedings of ECML 2008*.

Dekang Lin and Patrick Pantel. 2001. DIRT - discovery of inference rules from text. *Proc.Seventh ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young and Ralph Grishman. 2004. Annotating noun argument structure for NomBank. *Proc. LREC-2004*.

Adam Meyers, Michiko Kosaka, Heng Ji, Nianwen Xue, Mary Harper, Ang Sun, Wei Xu, and Shasha Liao. 2009. Transducing logical relations from automatic and manual GLARF. *Proc. Linguistic Annotation Workshop-III at ACL 2009.*

Bonan Min, Shuming Shi, Ralph Grishman, and Chin-Yew Lin. 2012. Ensemble semantics for large-scale unsupervised relation extraction. Microsoft Technique Report, MSR-TR-2012-51.

Frederick Newmeyer. 1980. *Linguistic Theory in America.* Academic Press, New York.

Hoifung Poon and Pedro Domingos. 2009. Unsupervised semantic parsing. *Proc. 2009 Conf. Empirical Methods in Natural Language Processing.*

Bryan Rink and Sanda Harabagiu. 2011. A generative model for unsupervised discovery of relations and argument classes from clinical texts. *Proc. 2011 Conf. Empirical Methods in Natural Language Processing.*

Naomi Sager, Carol Friedman, and Margaret S. Lyman. 1987. *Medical Language Processing: Computer Management of Narrative Data.* Addison-Wesley, Reading, MA.

Yusuke Shinyama and Satoshi Sekine. 2006. Preemptive Information Extraction using Unrestricted Relation Discovery. *Proc. NAACL 2006*.

Limin Yao, Aria Haghighi, Sebastian Riedel, Andrew McCallum. 2011. Structured Relation Discovery Using Generative Models. *Proc. EMNLP 2011*.

Alexander Yates and Oren Etzioni. 2007. Unsupervised Resolution of Objects and Relations on the Web. *Proc. HLT-NAACL 2007*.

Min Zhang, Jian Su, Danmei Wang, Guodong Zhou, and Chew Lim Tan. 2005. Discovering relations between named entities from a large raw corpus using tree similarity-based clustering. *Proc. Second Int'l Joint Conference on Natural Language Processing (IJCNLP).*