An Ontology-Based Opinion Mining Approach for the Financial Domain

Conference Paper · May 2012

DOI: 10.1007/978-3-662-46641-4_6

3 authors:

Juana María Ruiz-Martínez
University of Murcia
14 PUBLICATIONS **193** CITATIONS
SEE PROFILE

Rafael Valencia-García
University of Murcia
170 PUBLICATIONS **1,754** CITATIONS
SEE PROFILE

Francisco Garcia-Sanchez
University of Murcia
76 PUBLICATIONS **1,243** CITATIONS
SEE PROFILE

Adaptive Information Extraction and Sublanguage Analysis

Ralph Grishman
Computer Science Department
New York University
New York, NY 10003
grishman@cs.nyu.edu

1. Introduction[1]

Information extraction (IE) has made significant progress in the last decade. We have developed practical, efficient approaches to IE which have yielded modest levels of performance on general texts and quite good performance on restricted, 'semi-structured' texts. More notably, over the last few years there has been a blossoming of work in adaptive IE — the topic of this and other recent workshops — IE systems which can be rapidly and automatically (or semi-automatically) moved to new extraction tasks.

To date, these developments have been relatively little influenced by linguistic studies of the texts. In fact, the trend has been towards less linguistic analysis. Some early IE systems used full parsing and in a few cases relatively deep semantic analysis. Because of limitations of full parsing methods (particularly a decade ago) this gave way to a common methodology based on limited parsing and simple pattern matching. Adaptive IE systems have in many cases eschewed syntactic analysis entirely, and learned patterns based on pure word sequences (using regular expressions and Markov models) [Califf and Mooney 1999; Freitag and McCallum 1999]. Methods operating directly on the word sequence have in fact proven quite effective for 'semi-structured' texts, where the patterns of occurrence are quite repeatable and syntactic cues are quite minimal.

The systems which have been built have, in general, focused on plucking one or a few relations from a sea of facts. The choice of relations has generally been driven by practical demands or 'user needs'.

[1] The author's research is supported by the National Science Foundation under Grant IIS-0081962 and by the Defense Advanced Research Projects Agency under Grant N66001-00-1-8917 from the Space and Naval Warfare Systems Center San Diego. This paper does not necessarily reflect the position or the policy of the U. S. Government.

As we gain more confidence in creating these limited systems, we can set our sights higher and envision a system which extracts a broad range of facts from texts within a subject domain. These have been referred to as 'open domain extraction systems'. How should we go about building such systems? In particular, how should we go about identifying the facts to be extracted?

In addressing these questions, we can gain some guidance from linguistic studies of these issues … in particular, the studies of *sublanguage* and of *sublanguage information structures*.

2. Sublanguage

A sublanguage is the specialized form of a natural language which is used within a particular domain or subject matter [Grishman and Kittredge 1986]. Examples of sublanguages are the languages of weather reports, aircraft repair manuals, scientific articles about pharmacology, hospital radiology reports, and real estate advertisements. A sublanguage is characterized by a specialized vocabulary, semantic relationships, and in many cases specialized syntax.

Zellig Harris, focussing primarily on scientific and technical sublanguages, incorporated a formal notion of sublanguage into his mathematical theory of language [Harris 1968]. In this view, the sublanguage is a subset of the general language, and there are clear notions of sentences being unacceptable in the sublanguage, even if they are acceptable in the general language. For example, it would be acceptable in a biochemistry article to say "The polypeptides were washed in hydrochloric acid." but not "Hydrochloric acid was washed in polypeptides." The subset is closed under syntactic operations of the general language. This means that if a sentence is acceptable in the sublanguage, its syntactic transforms will be too (so if "The enzyme activated the process." is OK, so is "The process was activated by the enzyme.").

The subset which constitutes the sublanguage can be specified in terms of a set of sublanguage word classes and allowable combinations of these classes (just as the language is specified by syntactic word classes and their allowed combinations). These classes involve domain-specific entities such as (for biochemistry) cells and enzymes.

In this study of sublanguage, two issues are of particular interest to us as developers of adaptive IE systems: discovery methods and information formatting (structuring).

The idea of discovery methods was central to much of the work in structural linguistics. Levels of linguistic description were coupled with methods for discovering these descriptions (in principle) from text. In the case of sublanguage, through syntactic regularization and decomposition, one could reduce complex sentences to simple canonical syntactic structures (e.g., simple active declarative clauses; what Harris terms *kernel sentences*). These structures would then exhibit repetitive patterns of word choice, which could be captured in terms of sublanguage word classes and sublanguage sentence structures using these classes.

This analysis of the linguistic structure of texts in the domain also gives us insight into the *information structure* of the domain: what the basic classes of objects are in the domain, how they may be modified, and how they may be related to other objects. By combining the kernel sentence structures we have discovered, along with the associated modifiers and quantifiers, and possible connectives to other kernels, we can create a set of canonical structures for capturing the information in the sublanguage text. Harris directed several studies to develop such information structures for science sublanguages [Harris et al. 1989]. Sager has done such studies for scientific and medical texts, as well as developing information extraction procedures for filling such structures for some types of medical texts [Sager et al. 1987].

There are certainly differences between these sublanguage studies and open-domain IE. In the science sublanguage studies, the ideal is that of a sharply defined sublanguage, and the goal of information structuring is to be able to capture *all* the information in the sublanguage text. Open domain extraction seeks to identify the *most important relations* within a given text domain, where the boundaries of this domain (e.g., financial news) may not be very sharply defined. However, at their heart the goal is the same: to discover from a text the essential information relationships and the linguistic expressions of these relationships. We may hope, therefore, that each may learn from the methods developed by the other.

3. Methods

3.1. Pattern discovery

The methodology which has developed over the last few years for *unsupervised* (or lightly supervised) adaptive IE is particularly close to that for discovery of sublanguage kernel patterns. While most work on adaptive IE has used supervised training involving annotated corpora, several recent experiments have started from raw text.

In the area of extraction pattern discovery, the first such work was that of Riloff [1996]. She used a corpus which had been divided into documents which had been marked as relevant or not relevant to the extraction task. From the corpus (both relevant and non-relevant articles) she pulled out all word combinations in a predefined set of syntactic relations. She then compared the frequency of a combination in relevant texts to that in non-relevant texts. Those which occurred significantly more frequently in relevant texts were good candidates for extraction patterns.

This approach was extended by Yangarber et al. [2000a, 2000b]. In this recent work, the corpus is completely unmarked (not even for relevance). Discovery starts with a few word patterns which are known to be good indicators of the topic of interest. These patterns are used to retrieve some relevant documents; the relevant documents are used to identify one or more additional patterns (using a method similar to Riloff''s). This process repeats, expanding both the relevant document set and the pattern set. A related method from Sudo et al. [2001], applied to Japanese, uses information retrieval methods to obtain relevant documents and sentences, and then builds patterns from them.

In Riloff's case, the documents are those specifically judged relevant to an extraction task. In Yangarber's the documents are those related to a 'topic' description, stated in terms of patterns. It seems natural to extend the same approach to a broader collection, covering an entire domain. The patterns retrieved will then not be expected to cover a single event type, but rather a range of events — the most common event types for the domain. This would be similar to earlier experiments aimed at gathering sublanguage kernel patterns automatically from parsed text [Grishman et al. 1986, Sager 1986].

3.2. Building information structures

In the manual analyses of sublanguage information structures, in order to minimize the number of different patterns, syntactic regularization was applied as much as possible to combine patterns. This included simple things like reducing all clause forms (active, passive, relative, reduced relative, etc.) to a single form. It also involved combining verbal and nominalized forms.

Because of the limitations of high-performance computer parsers, less (or no) regularization was performed in the automatic experiments. Riloff's work involved a fixed set of separate syntactic relations; Yangarber's involved some regularization of clause structures. Some of the sublanguage analyzers employed more extensive regularization, although not for nominalizations. More extensive automatic syntactic regularization would clearly be a benefit if we are seeking to collapse related structures.

In addition, we would want to use lexical semantics (at least, synonymy or near-synonymy) to identify related patterns. Again, this was done by hand in the sublanguage analysis, but one could imagine using lexical semantics resources as part of IE pattern discovery (although, I believe, this has not been done until now).

The basic sublanguage kernels combine in only limited ways, and so (as we noted earlier) they can be assembled into larger information structures — combining a kernel with modifiers of the verb, modifiers of its arguments, and sometimes other kernels with which it is regularly connected. These larger structures are then much more useful for manual inspection and fact retrieval. We see the same process going on, implicitly, when people design IE templates, often combining several types of relationships into a single template. This may have to be done manually but could receive some guidance from distributional analysis.

3.3. Word class discovery

Both the sublanguage experiments and adaptive extraction have used some source of word class information. For sublanguage discovery experiments, a sublanguage lexicon listed the main sublanguage classes and their members; for some of the experiments of extraction pattern discovery from unannotated text, the *names* were classified using a named entity tagger (into people, organizations, job titles, locations, etc.). Some such (partial) source of word class information can significantly enhance pattern discovery from unannotated text.

Word class discovery (or extension) has itself been the object of many experiments. Riloff and Jones (1999) develop the classes needed for an extraction task by assembling a set of contexts for each class. Yangarber (2000) also discovers classes as a by-product of pattern discovery. This work is similar to a long history of experiments on finding sublanguage word classes using distributional experiments (e.g., [Hirschman et al 75]).

The experience of sublanguage analysis is that the pattern creation and the word class creation must go hand-in-hand. We may expect, therefore, that these two processes will be more tightly coupled in future adaptive IE systems.

4. Prospects

The idea of 'open domain' adaptive information extraction is very appealing. It holds out the hope that for a new domain, we would be able to automatically identify the primary relationships, and then create systems which can extract these relationships. This would provide data bases which users could browse, use to retrieve documents, or use to answer questions. As I have tried to point out, the challenge is similar to that which linguists have been considering for some time, to identify the information structures for a new sublanguage.

Our experience with both sublanguage discovery and (more recently) with adaptive IE tells us that the task will not be easy. Getting the highest-frequency patterns is not too difficult, but to go further we need high-quality linguistic analyzers. We will need good (and broader coverage) named entity analysis to find name classes. We will need good parsers; while semi-structured text can be analyzed at the level of word sequences, that will probably not be possible for complex scientific, technical, or commercial sublanguages. In particular, we will want parsers which can regularize their input, and do so reliably. Without all of this, our linguistic data may be too noisy for the analysis we wish to perform.

In addition, we will want more sophisticated distributional analysis: to couple pattern and word class discovery, and to build structures above the kernel level.

Many of these ingredients are now coming together; we see, for example, steady progress in corpus-trained parsing and in adaptive named entity analysis. The next decade of adaptive IE promises to be exciting.

References

[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.

[Freitag and McCallum 1999] Dayne Freitag and Andrew McCallum, Information Extraction with HMMs and Shrinkage. *Proc. Workshop on Machine Learning and Information Extraction, AAAI-99*, 1999.

[Grishman and Kittredge 1986] Ralph Grishman and Richard Kittredge, editors. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Lawrence Erlbaum Assoc., Hillsdale, NJ, 1986.

[Grishman et al. 1986] Ralph Grishman, Lynette Hirschman, and N. T. Nhan. Discovery Procedures for Sublanguage Selectional Patterns: Initial Experiments. *Computational Linguistics*. **B 12** (3), 205-216, 1986.

[Harris 1968] Zellig Harris. *Mathematical Structures of Language*. Wiley-Interscience, New York, 1968.

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

[Hirschman et al. 1975] Lynette Hirschman, Ralph Grishman, and Naomi Sager. Grammatically-based Automatic Word Class Formation. *Information Processing and Management* **11**, 39 (1975).

[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)*.

[Sager 1986] Naomi Sager. Sublanguage: Linguistic Phenomenon, Computational Tool. In [Grishman and Kittredge 1986].

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

[Sudo et al. 2001] Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman. Automatic Pattern Acquisition for Japanese Information Extraction. *Proc. HLT 2001*, San Diego, CA, 2001.

[Yangarber et al. 2000a] Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. Unsupervised Discovery of Scenario-Level Patterns for Information Extraction. *Proc. Sixth Applied Natural Language Processing Conf.*, Seattle, WA, April-May, 2000, 282-289.

[Yangarber et al. 2000b] 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.