Large-scale Controlled Vocabulary Indexing for Named Entities

Mark Wasson
LEXIS-NEXIS, a Division of Reed Elsevier plc
9443 Springboro Pike
Miamisburg, Ohio 45342 USA
mark.wasson@lexis-nexis.com

Abstract

A large-scale controlled vocabulary indexing system is described. The system currently covers almost 70,000 named entity topics, and applies to documents from thousands of news publications. Topic definitions are built through substantially automated knowledge engineering.

1 Introduction

The goal of the Entity Indexing R&D program at LEXIS-NEXIS is to add controlled vocabulary indexing for named entities to searchable fields in appropriate news documents across thousands of news publications, where documents include both incoming news articles as well as news articles already in the LEXIS-NEXIS archives. A *controlled vocabulary term* (*CVT*) is a consistently specified topic indicator that users can incorporate into their queries in order to retrieve documents about the corresponding topic. When a CVT is added to an appropriate field in a document, it can be included in a Boolean query using

field-name(controlled vocabulary term)

The initial Entity Indexing release focused on companies as topics. For company indexing, the *primary CVT* is a standard form of the company name. When we add a company CVT to a document, we often also want to add *secondary CVTs* to the document that specify attributes of that company. Attributes may include the ticker symbol, SIC product codes, industry codes and company headquarters information. Secondary CVTs allow customers to easily search on groups of companies that have one or more attributes in common, such as searching for documents about banks in our set of companies that are headquartered in Utah.

It is generally easy to get high recall with Boolean queries when searching for documents about named entities. Typically the query will only need a short form of the entity's name. For example, the query

American

will retrieve virtually every document that mentions American Airlines. Of course, this query results in poor precision due to the ambiguity of *American*. The problem we wanted to address with controlled vocabulary indexing is to help online customers limit their search results to only those documents that contain a *major reference* to the topic, that is, to documents that are substantially about the topic.

Because of the volume of news data we have, it is necessary that we fully automate the document categorization and indexing step in our data preparation process. LEXIS-NEXIS adds 100,000 news articles daily to its collection of over 2 billion documents.

For marketing and product positioning reasons, we want to provide indexing for tens of thousands of companies, where companies are targeted based on their presence on the New York, American and NASDAQ exchanges or on revenue-based criteria. Although such selection criteria help us explain the product feature to customers, it does not ensure that the targeted companies actually appear all that often in the news. In fact, for many targeted companies there is little training and test data available.

Our company indexing system should address the following business product requirements:

- Assign primary and corresponding secondary CVTs to appropriate documents

- Add CVTs only to those documents that contain major references to the topic(s)
- Process documents fast; target 30,000 characters per CPU second
- Apply tens of thousands of topics to the data in a single pass
- Minimize the cost of developing and maintaining topic definitions

Also, we target 90% recall and 95% precision when using the CVTs to retrieve major reference documents about the corresponding companies.

2 Related Work

The Carnegie Group's Text Categorization Shell (TCS) (Hayes, 1992) uses shallow knowledge engineering techniques to categorize documents with respect to large sets of predefined topics. Each topic requires the development of a rule set that includes terms, contextual information, weighting, if-then rules and other pattern matching operations. This initially involved a manual, iterative approach to rule development, although Hayes (1992) discusses their intent to explore ways to automate this. TCS accuracy is quite good. One application deployed at Reuters achieved 94% recall and 84% precision. Other reported tests achieved recall and precision rates of 90% or better.

SRA's NameTag (Krupka, 1995) uses a pattern recognition process that combines a rule base with lexicons to identify and extract targeted token classes in text, such as company, people and place names. It achieved 96% recall and 97% precision when tested on Wall Street Journal news articles at MUC-6. Aone et al. (1997) describes a NameTag-based application for indexing English and Japanese language texts. NameTag addresses a key weakness of approaches that use predefined topic sets. Predefined topic sets are inherently limited in their coverage to those topics that have been explicitly defined. NameTag can recognize any number of companies or entities of other domains whose names have structure that the rules can recognize. (Not all entity domains have structure that pattern recognition processes can exploit, e.g., product names).

A problem for pattern recognition approaches has to do with our requirement to assign CVTs. Pattern recognition approaches extract patterns such as company names as they appear in the text. Limited coreference resolution may link variant forms of names with one another to support choosing the best variant as a "semi-controlled" vocabulary term, but this does not allow for the assignment of true primary and secondary CVTs. SRA has attempted to address this through its Name Resolver function, which reconciles extracted names with an authority file, but the authority file also limits scope of coverage for CVTs to those that are defined and maintained in the authority file. The system must also go beyond straight recognition in order to make a distinction between documents with major references to the targeted entities and documents with lesser or passing references. SRA's NameTag addresses this with the calculation of relevance scores for each set of linked variant forms.

Preliminary research suggests that recognizing named entities in data and queries may lead to a significant improvement in retrieval quality (Thompson & Dozier, 1999). Such an approach may complement Entity Indexing, but it does not yet meet the controlled vocabulary indexing and accuracy requirements for Entity Indexing.

Our own Term-based Topic Identification (TTI) system (Leigh, 1991) combines knowledge engineering with limited learning in support of document categorization and indexing by CVTs. We have used TTI since 1990 to support a number of topically-related news or legal document collections. Categories are defined through *topic definitions*. A definition includes terms to look up, term weights, term frequency thresholds, document selection scoring thresholds, one or more CVTs, and source-specific document structure information. Although creating TTI topic definitions is primarily an iterative, manual task, limited regression-based supervised learning tools are available to help identify functionally redundant terms, and to suggest term weights, frequency thresholds and scoring thresholds. When these tools have been used in building topic definitions, recall and precision have topped 90% in almost all tests.

3 Approach

TTI was originally proposed as a tool to support controlled vocabulary indexing, and most early tests focused on narrowly defined legal and news topics such as *insurable interest* and *earthquakes*. TTI was also tested on a number of companies, people, organizations and places as topics. TTI was first put into production to categorize documents by broadly-defined topics such as *Europe political and business news* and *federal tax law*.

When we began investigating the possibility of creating Entity Indexing, TTI was a natural starting point. It had demonstrated high accuracy and flexibility across a variety of topics and data types. Three problems were also apparent. First, TTI would not scale to support several thousand topics. Second, it took a long time to build a topic definition, about one staff day each. Third, topics were defined on a publication-specific basis. With then-700 publications in our news archives in combination with our scale goals and the time needed to build topic definitions, the definition building costs were too high. We needed to scale the technology, and we needed to substantially automate the topic definition-building process.

For Entity Indexing, we addressed scale concerns through software tuning, substantially improved memory management, a more efficient hash function in the lookup algorithm, and moving domain-specific functionality from topic definitions into the software. The rest of this paper focuses on the cost of building the definitions.

3.1 Analyzing Companies in the News

In order to reduce definition building costs, we originally believed that we would focus on increasing our reliance on TTI's training tools. Training data would have to include documents from a variety of publications if we were to be able to limit definitions to one per topic regardless of the publications covered.

Unfortunately the data did not cooperate. Using a list of all companies on the major U.S. stock exchanges, we randomly selected 89 companies for investigation. Boolean searches were used to retrieve documents that mentioned those companies. We found that several of these companies were rarely mentioned in the news. One company was not mentioned at all in our news archives, a second one was mentioned only once, and twelve were mentioned only in passing. Several appeared as major references in only a few documents. In a second investigation involving 40,000 companies from various sources, fully half appeared in zero or only one news document in one two-year window. We questioned whether we even wanted to create topic definitions for such rarely occurring companies. Again, marketing and product positioning reasons dictated that we do so: it is easier to tell customers that the product feature covers *all* companies that meet one of a few criteria than it is to give customers a list of the individual companies covered. It is also reasonable to assume that public and larger companies may appear in the news at some future point.

3.2 Company Name Usage

While analyzing news articles about these companies, we noted how company names and their variants were used. For most companies discussed in documents that contained major references to the company, some form of the full company name typically appears in the leading text. Shorter variants typically are used for subsequent mentions as well as in the headline. Corresponding ticker symbols often appear in the headline or leading text, but only after a variant of the company name appears. In some publications, ticker symbols are used as shorter variants throughout the document. Acronyms are somewhat rare; when they are used, they behave like other shorter variants of the name.

Shorter variants typically are substrings of the full company name beginning with the leftmost words in the name. For a company named

Samson Computing Supply Inc.

shorter variants might include

Samson Computing Supply
Samson Computing
Samson

Acronyms typically exist only for companies whose names consist of at least two words in addition to company designators such as *Inc.* or *Corp.* There is no consistency as to whether the company

designators contribute to acronyms. Thus for the above example

SCSI
SCS

are both potential acronyms.

Generating such variants from a full form of a company name is straightforward. Similarly, rules working with a table of equivalences can handle common abbreviations, so the variant

Samson Computing Supply Incorporated

can be generated from

Samson Computing Supply Inc.

We assign weights to term variants based on term length and the presence or absence of company designators. Longer variants with designators are regarded as less ambiguous than shorter variants without designators, and thus have higher weights. A table of particularly problematic one-word variants, such as *American*, *National* and *General*, is used to make a weighting distinction between these and other one-word variants. One-word variants are also marked so they do not match lower case strings during lookup. A label is assigned to each variant to indicate its function and relative strength in the document categorization process.

3.3 Generating Topic Definitions

Our company controlled vocabulary indexing process requires definition builders to provide a primary CVT and zero or more secondary CVTs for each targeted company. The CVTs are the primary input to the automatic topic definition generation process. If the definition builder provides the company name and ticker symbol to be used as CVTs for some company, as in

#CO = Samson Computing Supply Inc.
#TS = SMCS (NYSE)

the following definition can be generated automatically:

#NAME1 = Samson Computing Supply Inc.
#NAME1 = Samson Computing Supply Incorporated
#NAME2 = Samson Computing Supply
#NAME3 = Samson Computing
#NAME4 = Samson
#TS @= SMCS {upper case only}
#ACRONYM @= SCSI {upper case only}
#ACRONYM @= SCS {upper case only}
#BLOCK @= samson {do not match lower case}

That two acronyms were generated points out a potential problem with using robust variant generation as a means to automatically build topic definitions. Overgeneration will produce some name variants that have nothing to do with the company. However, although overgeneration of variants routinely occurs, testing showed that such overgeneration has little adverse effect.

This approach to automatically generating topic definitions is successful for most companies, including those that appear rarely in our data, because most company names and their variants have consistent structure and use patterns. There are exceptions. Some companies are so well-known that they often appear in news articles without the corresponding full company name. Large companies and companies with high visibility (e.g., Microsoft, AT&T, NBC and Kmart) are among these. Other companies simply have unusual names. Our authority file is an editable text file where definition builders not only store and maintain the primary and secondary CVTs for each company, but it also allows builders to specify exception information that can be used to override any or all of the results of the automatic definition generation process. In addition, builders can use two additional labels to identify especially strong name variants (e.g., *IBM* for International Business Machines) and related terms whose presence in a document provide disambiguating context (e.g., *Delta*, *airport* and *flights* for American Airlines, often referred to only as *American*). For our initial release of 15,000 companies, 17% of the definitions had some manual intervention beyond providing primary and secondary CVTs. Entity definitions built entirely manually usually took less than thirty minutes apiece. Overall, on average less than five minutes were spent per topic on definition building. This includes the time used to identify the targeted companies and add their primary and secondary CVTs to the authority file. Populating the authority

file is required regardless of the technical approach used.

3.4 Applying Definitions to Documents

All topic definitions contain a set of labeled terms to look up. The document categorization process combines these into a large lookup table. A lookup step applies the table to a document and records term frequency information. If a match occurs in the headline or leading text, extra "frequency" is recorded in order to place extra emphasis on lookup matches in those parts of the document. If the same term is in several definitions (e.g., *American* is a short name variant in hundreds of definitions), frequency information is recorded for each definition.

Once the end of the document is reached, frequency and term label-based weights are used to calculate a score for each topic. If the score exceeds some threshold, corresponding CVTs are added to the document. Typically a few matches of high-weight terms or a variety of lower-weighted terms are necessary to produce scores above the threshold. A document may be about more than one targeted topic.

3.5 System Implementation

The tools used to build and maintain topic definitions were implemented in C/C++ on UNIX-based workstations. The document categorization process was implemented in PL1 and a proprietary lexical scanner, and operates in a mainframe MVS environment.

4 Evaluation

In the final pre-release test, Entity Indexing was applied to more than 13,500 documents from 250 publications. Each document in the test was reviewed by a data analyst. Several of these were also reviewed by a researcher to verify the analysts' consistency with the formal evaluation criteria. Recall was 92.0% and precision was 96.5% when targeting documents with major references. Additional spot tests were done after the process was applied in production to archived documents and to incoming documents. These tests routinely showed recall and precision to be in the 90% to 96% range on over 100,000 documents examined.

Some recall errors were due to company names with unusual structure. Many such problems can be addressed through manual intervention in the topic definitions. Some publication styles also led to recall errors. One publisher introduces a variety of unanticipated abbreviations, such as *Intnl* for *International*. Trade publications tend to use only short forms of company names even for lesser known companies. Those companies may be well-known only within an industry and thus to the audience of the trade publication. These types of problems can be addressed through manual intervention in the topic definitions, although for the abbreviations problem this is little more than patching.

Capitalized and all upper case text in headlines and section headings was a routine source of precision errors. These often led to unwanted term matching, particularly affecting acronyms and one-word company name variants. Different companies with similar names also led to precision problems. This was particularly true for subsidiaries of the same company whose names differed only by geographically-distinct company designators.

5 Discussion

Entity Indexing applies in production to tens of thousands of documents daily from thousands of news publications. Further tests have shown that we can reach comparably high levels of accuracy for company topics when Entity Indexing is applied to financial, patents and public records sources. Accuracy rates dropped about 10% in tests applied to case law parties.

Since the initial completion of the company indexing work, Entity Indexing has been extended to cover more companies and other named entities, including people, organizations and places. Topic definitions have been built for almost 70,000 entities, including over 30,000 companies. Entity Indexing applies these definitions during a single pass of the data, processing more than 86,000 characters (approximately 16 news documents) per CPU second.

The approach used for the other named entity types is similar to that for companies. However, because most of the place names we targeted lacked useful

internal structure, manual intervention was a part of creating all 800 definitions for places. Accuracy rates for people, organization and geographic indexing are comparable to those for company indexing.

Knowledge engineering can be a bottleneck in building large-scale applications, which is why machine learning-based approaches are often preferred, but there has been little work in quantifying the difference between the two approaches in linguistic tasks (Brill & Ngai, 1999). In our case, adopting a machine learning-based approach was a problem not just because we lacked annotated training data, but for many of the topics we were required to target we had little or no available data at all. However, because of the regularity we observed in company name variants and their use across a variety of news sources, we determined that the knowledge engineering task would be quite repetitive and thus could be automated for most companies.

Our Entity Indexing system meets all of our business requirements for accuracy, scale and performance. We have also substantially automated the definition building process. Even if a number of documents were available for each targeted entity, it is unlikely that we would see the time needed to populate the authority file and to create and annotate appropriate training data in a machine learning-based approach fall much below the under five minutes we average per definition with our chosen approach.

Acknowledgments

I would like to thank colleagues Mary Jane Battle, Ellen Hamilton, Tom Kresin, Sharon Leigh, Mark Shewhart, Chris White, Christi Wilson and others for their roles in the research, development and ongoing production support of Entity Indexing.

References

Aone, C., Charocopos, N., & Gorlinsky, J. (1997). *An Intelligent Multilingual Information Browsing and Retrieval System Using Information Extraction*. Proceedings of the Fifth Conference on Applied Natural Language Processing.

Brill, E., & Ngai, G. (1999). *Man vs. Machine: A Case Study in Base Noun Phrase Learning*. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.

Hayes, P. (1992). *Intelligent High-volume Text Processing Using Shallow, Domain-specific Techniques*. In P. Jacobs (ed.), Text-based Intelligent Systems. Lawrence Erlbaum.

Krupka, G. (1995). *SRA: Description of the SRA System as Used for MUC-6*. Proceedings of the Sixth Message Understanding Conference (MUC-6).

Leigh, S. (1991). *The Use of Natural Language Processing in the Development of Topic Specific Databases*. Proceedings of the Twelfth National Online Meeting.

Thompson, P., & Dozier, C. (1999). *Name Recognition and Retrieval Performance*. In T. Strzalkowski (ed.), Natural Language Information Retrieval. Kluwer Academic.