Text Corpora, Local Grammars and Prediction

Hayssam Traboulsi, David Cheng, Khurshid Ahmad

Text Corpora, Local Grammars and Prediction

Hayssam Traboulsi, David Cheng and Khurshid Ahmad

Department of Computing, University of Surrey, Guildford, Surrey, GU2 7XH, UK
{h.traboulsi, d.cheng, k.ahmad}@surrey.ac.uk

Abstract

We present a corpus-based method for identifying and learning patterns describing *events* in a specific domain by examining the manner in which: (a) a small number of keywords in the domain are distributed throughout the corpus; and, (b) a local grammar that is idiosyncratic of a class of events in the domain, governs the usage of the keywords. We tested our method against a corpus of 3.63 million words. The results show promise. More importantly, the method can be applied to any arbitrary domains.

1. Introduction

Work on extracting information automatically from news that reports about an *event* – a significant occurrence located at a single point in space-time - typically focuses on filling templates of at least three kinds: First, Template Element (TE) for representing attributes/values of key entities inserted in a given *event* - people, places, organisations, objects and concepts; Second, Template Relation (TR) for encoding n-ary relations between the entities; Third, Scenario Template (ST) for detecting/representing relations between entities in a particular type of *events*.

Typically, information extraction (IE) systems appear to have a high precision (85%) for TE and TR filling but a lower precision for ST (65%). The recall statistic is good for TE (86%) but is degraded for TR (61%) and gets worse for ST (42%) (Humphreys 2000). The process of populating such templates requires sets of assorted patterns to be collected and made available to the core IE engine. These patterns commonly consist of regular expressions and associated mappings from syntactic to logical form. In addition, these patterns need to be revised every time a new set of events in a specific domain is identified to be extracted. The construction of a pattern base for a new domain is recognized as a lengthy and costly process – one of the principal obstructions to the adaptation of IE technology to new domains (Yangarber & Grishman, 1997).

This paper describes a corpus-based method for identifying and learning patterns describing events in a specific domain. For instance, in the financial trading domain, one may be interested in the changes in the value of inanimate entities, e.g. currencies, derivatives and shares. The change, actual or perceived, in the values of these instruments is communicated in as unambiguous a manner as is possible: Keywords reduce ambiguity and what appeared to be structured in a local grammar are used to inter-relate keywords.

Local grammar movement was started by Harris (1991) to discuss recursive noun phrases that are commonly found in specialist literature like biochemistry. Gross (1993) extended the application of local grammars for articulating date, times and for financial reports. Sinclair and Barnbrook (1995) suggested that dictionary definitions are formatted according to local grammar patterns. Key-sun and Jee-Sun (1997) used much the same approach to Korean proper nouns. Ranchhod (1999) has extended Gross' work into Portuguese. Mason (2002) has used local grammars 'discovering' verb patterns in English.

We describe a corpus-based method for identifying patterns by examining the manner in which: (a) a small number of keywords in the domain are distributed throughout the corpus; and, (b) a local grammar that is idiosyncratic of a class of events in the domain, governs the usage of the keywords. The method includes three steps: frequency analysis, collocation analysis and concordance analysis. The execution of the methods leads to a local grammar specific to the domain (Section 2). Next we describe our corpus and demonstrate how our method can be used to develop local grammars (Section 3). The conclusion and future work are presented in Section 4.

2. Method

Documents written in English comprise two major linguistic units: the high frequency closed class words, determiners, prepositions, conjunctions, some verbs and modal verbs; and lower frequency open class words such as nouns, adjectives and adverbs used in technical or specialist languages. In general, the closed class words are indicative of the natural language being used and the open class words indicate the topic of discussion. In everyday language, fewer open class words are used than in special language texts. Specialists use distinct open class words with much higher frequency.

Local grammars are rules that govern the simultaneous choice of a set of words used in a specialist context. For example, sentences/clauses used for telling time and dates are one of such simultaneous choice of words. Others include what appear to be telegraphic statements in specialist domains. Let the sentences in a local grammar are:

$$S := P_1 + P_{2+} P_3$$

Where P_i are i patterns. Table 1 shows some typical patterns.

Domain	$P_1 + P_2 + P_3$	Authors
Biochemistry	The ***polypeptides*** were ***washed*** in ***acid*** The ***protein*** was ***treated*** with ***acid***.	Harris 1991
Transport	***accidents*** including ***cars*** rose/fell from # to # ***accidents*** including ***ships*** rose/fell from # to #	BNC Online
Finance	***The Dow Jones fell 14.50*** points. ***Dow Jones rose 15*** points.	Gross 1997

Table 1: Local grammar for typical patterns.

Gross (1997) has discussed local grammars for the financial domain to suggest some local grammars for

sentences that are used to indicate the status of a financial instrument. It appears that most local grammars are constructed intuitively. Our work is an extension of Gross in that we suggest that a corpus of texts can be used systematically to find such grammars. The local grammars deal with a large number of open class words – nouns, verbs, adjectives - and restricted closed words such as prepositions, numbers and punctuation marks. The key point, at least for us, is that the open class words are frequently used in specialist domains. And, it is this simultaneous use of the frequent open class words with other well-defined words, which might lead to meaning bearing sentences that are governed by a local grammar. Typically, such meaning bearing sentences are attributed to people or organisations of reputation and authority. This identification of a proper noun related to reputation and authority will enhance the value of meaning bearing sentences.

A related notion to that of local grammar is that of collocation. Collocation is related to the localist idiom principle, as opposed to the open choice principle governed by (universal) grammaticalness (Sinclair, 1991: 109-115). There are frequently used words and less frequent words that are simultaneously used. For example, the frequently used "washed" may occur with many proteins and acids in Table 1. Many financial instruments, for example, currencies and indices co-occur with the frequently used verbs *fell* and *rose*. In addition *said* can be used with a number of people and organization names. This, collocation between high frequency words and many lower frequency words is called "downward collocation" (ibid: 115). Generally, collocation strength is computed by the so-called "collocation strength" of a word pair. The pair may be neighbours, or may co-occur with other interspersing words: up to five left neighbours and five right neighbours of a high frequency word appear to be significant. Smajda (1994) defines the strength k_i of the collocates w_i of the high frequency word *w* as

$$k_i = \frac{freq_i - \overline{f}}{\sigma}$$

where $freq_i$ is the frequency of the collocation of w_i with *w*; $\overline{f}$ is the average frequency and σ is the standard deviation.

We present a systematic method of finding such collocation between the frequently occurring words in a text corpus and other lower/higher frequency words. The collocation between non-neighbouring collocates may be just as useful between neighbouring collocates. The strength of their collocations can be statistically tested under certain statistical assumptions (Smajda 1994).

Starting from a corpus of specialist text, most frequently used words are selected and their collocations are identified. A set of concordances are produced and visually examined. This examination leads to a local grammar. The local grammar is used either for building a linguistic description of the language governing specialist texts or for purposes of recognising patterns of local grammar usage in unseen texts. Such recognition is useful for information extraction tasks. Furthermore, using the above procedure for identifying such patterns and adding them to the local grammar base can incorporate unrecognised patterns. This updating of the local grammars using unseen texts is an example of the adaptation or perhaps more formally that of adaptive learning (see Figure 1).

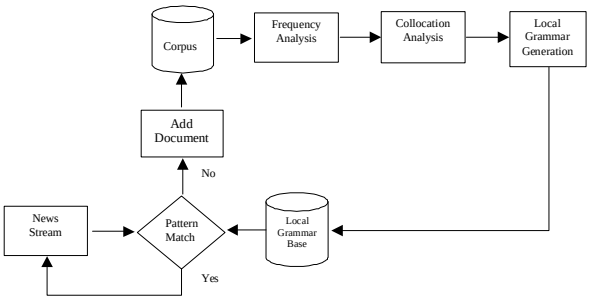

Figure 1: Analysis stages to discover patterns.

3. Data and Analysis

3.1 Data

We have used Reuters Financial News Service as the basis of our corpus. Our Reuters Corpus comprises 9,063 XML-tagged texts comprising 3.63 Million tokens published during Jan – Dec 2002. The average length of the news reports is 400 tokens. The subjects typically covered by Reuters Financial News Service are: Company Outlooks, Company Results, Economic Indicators, Funds and Initial Public Offering News.

3.2 Frequency Analysis

A frequency analysis of the corpus shows that the frequency of the first 50 most frequent tokens accounts for 40% (1.45 Million tokens). The first 10 most frequent tokens includes, in addition to some closed class words, key open class words like *said* (43643) and *percent* (41083) (see Table 2).

Rank	Token	Cumulative Relative Frequency
1-10	the, to, of, in, a, and, *s*, **said**, on, **percent**	21.70%
11-20	for, it, its, that, by, at, was, is, as, with	7.30%
21-30	from, **year**, has, but, be, **million**, which, after, have, **market**	4.44%
31-40	would, **up**, **pounds, shares**, will, are, **Reuters**, had, this, an	3.37%
41-50	were, **London**, not, u, he, we, **company, group**, last, **billion**	2.71%

Table 2: Distribution of the first 50 most frequent tokens in Reuters' corpus.

3.3 Collocation Analysis

We now attempt to find the collocation of the two most frequent open class words *said* (43643) and *percent* (41083), in our corpus. It is our intuitive expectation that *percent* will collocate with numbers, or changes in the value of a number, and *said* with a proper noun.

We have computed the downward collocation of *said* and *percent*, and selected five strong collocates for each of the two words (see Table 3a & 3b).

	f	Left	Right	Total	K-score
said	43643				
analyst	2844	672	1568	2240	8.21
after	15454	1466	211	1677	6.12
analysts	4907	795	583	1378	5.7
but	16859	963	85	1048	3.79
chairman	2101	155	90	245	2.14

Table3a: Collocates of *said* (f=43643).

	f	Left	Right	Total	K-score
percent	41083				
up	13465	3916	557	4473	16.97
down	7971	3319	482	3801	14.6
year	22008	1002	2869	3871	14.41
rose	4257	2653	187	2840	9.7
fell	4558	2504	209	2713	9.39

Table 3b: Collocates of *percent* (f=41083).

A closer examination of *said* collocations revealed that *said* collocates with tokens such as the definite and indefinite articles (*the*, *a*) and the speech mark ("), which have a higher frequency as compared to *said* itself (Table 3c). This upward collocation, as we show later, is quite important and cannot be ignored.

	f	Left	Right	Total	K-score
said	43643				
the	217426	8459	10613	19072	70.59
“	63804	14990	24	15014	55.55
a	86544	2312	5914	8226	30.39

Table 3c: Collocates of *said* (f=43643).

3.4 Concordance Analysis and Local Grammar Generation

The concordance analysis for *said* showed the dominance of patterns that have structures similar to those illustrated in Table 4a.

said	*Thomas Deitz, media analyst at Merrill Lynch*	**.**
said	*Aberdeen Asset Management fund manager James Laing*	**.**
said	*Chief Executive Ian Russell*	**.**
“[…]”	*Princess Chief Executive Peter Ratcliffe*	**said**
The	*Federal Reserve Bank of Philadelphia*	**said**
a	*YES spokesman*	**said**

Table 4a: Concordances of *said*.

We note that in addition to referring to named persons, *said* is also used simultaneously with named organisations We refer to these expressions, as named entities (NE). The structures of most frequent NE variants are shown in Table 4b.

Pattern	No. of patterns	Variant	No. of Variants
“[…]**”** ***said*** *NE.*	5739	**“**[…]**”** ***said*** *PN, TITLE at ORG.*	1319 (23%)
		“[…]**”** ***said*** *ORG TITLE PN.*	345 (6%)
		“[…]**”** ***said*** *ORG/PN.*	287 (5%)
		10 other Variants (<5% each)	3788 (66%)
“[...]**”** *NE* ***said***	3440	*"*[…]*"* *ORG/PN* ***said***	1995 (58%)
		"[…]*"* *TITLE PN* ***said***	379 (11%)
		"[…]*"* *ORG TITLE PN* ***said***	344 (10%)
		4 other Variants (<5% each)	722 (21%)
the *NE* ***said***	1368	***the*** *ORG* ***said***	1368 (100%)
a *NE* ***said***	276	***a*** *ORG TITLE* ***said***	127 (46%)
		a *TITLE for ORG* ***said***	36 (13%)
		a *TITLE for the ORG* ***said***	16 (6%)
		7 other Variants (<5% each)	97 (35%)

Table 4b: The structure of the most frequent of *said* patterns. *PN*: person name.*ORG*: organisation name.*MOD*: Modifier used before the organisation name (e.g. banker) and *TITLE*: the professional title of the speech originator.

The simultaneous choice of the speech mark and *said* as proximate neighbours is most widely used (5739). The next is *said* and the speech mark intercepted by a NE (3440). Less frequent is the choice of *said* and the closed class words *the* (1368) and *a* (276) respectively. From the analysis of Table 4b, we obtain a cascade of Finite State Automata (FSA) for *said* patterns, as shown in Figure 2a.

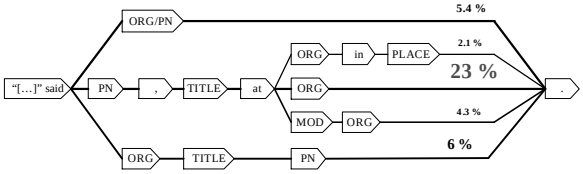

Figure 2a: Cascade of FSA representing some of the most frequent variants of *“[…]” said NE.*

The concordance analysis of *percent* patterns showed high occurrence of expressions enclosing cardinal numbers denoting changes in the price of financial instruments in these patterns. We refer to these numbers, as X. Examples of *percent* concordances are shown in Table 5a.

shares **rose**	*2.83*	**percent** to 584 pence
Payments , **rose**	*by 0.1*	**percent** on the month
food sales **up**	*14*	**percent** but bar sales flat
showing prices **up**	*nearly five*	**percent** on the month
airline easyjet **fell**	*9.7*	**percent** , with investors
public properties, **fell**	*as much as 24*	**percent** on news

Table 5a: Concordances of *percent*.

Similarly, the context of each of *percent* patterns was investigated.

Pattern	No. of patterns	Variant	No. of Variants
rose [..] *%*	2554	rose X %	1372 (54%)
		rose by X %	74 (2%)
		27 other Variants (<1% each)	1108 (44%)
fell [..] *%*	2423	fell X %	1134 (47%)
		fell by X %	51 (3%)
		34 other Variants (<1% each)	1238 (50%)
down [..] *%*	3319	down X %	1119 (34%)
		down from X %	43 (1%)
		66 other Variants (<1% each)	2157 (65%)
up [..] *%*	3769	up X %	822 (22%)
		up by X %	86 (2%)
		76 other Variants (<1% each)	2861 (76%)

Table 5b: The structure of *percent* patterns.

The simultaneous choice of the movement words - *rose*, *fell*, *up*, *down* – plus X and *percent* as proximate neighbours are the most dominant ones. Variants with frequency less than 1% are omitted from Table 5b. Analysis of the results above (Table 5b) helps us to construct the following cascade of *rose* FSA (Figure 2b):

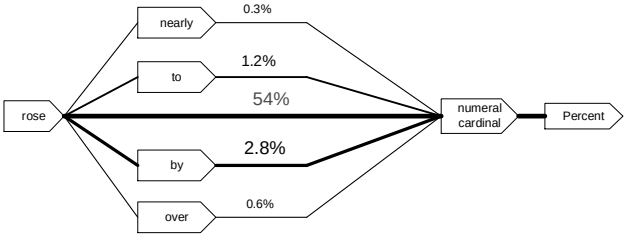

Figure 2b: Cascade of FSA representing some of the most frequent variants of rose patterns.

An extended study of the contexts of these patterns showed that they are mostly used to indicate upward and downward movements (changes) in the share price of some company (e.g. British American Tobacco rose 3.9 percent). Table 3 shows the distribution of *rose* extended patterns across our corpus. Similar behaviour was observed for *fell*, *up*, and *down*.

Collocate	Jan - Oct	Nov	Dec	Average	σ
$\mathbf{f_{percent}}$	34141	4264	2678	3370	749
rose X %	1140	130	111	115	27
$\mathbf{f_{said}}$	35810	4630	3202	3637	668
said Y	4681	623	435	478	106
Y rose X %	763	60	64	74	24

Table 3: The distribution of *rose* most frequent patterns across the whole corpus, σ is the standard deviation.

A cascade of FSA illustrating the local grammars of market sentiments is revealed in Figure 3. The numeral cardinal is the value by which the share price of some company went up or down.

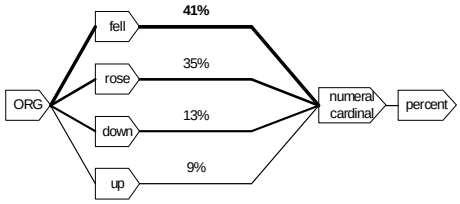

Figure 3: The FSA of *percent* most frequent patterns.

4. Afterword

The discussion in this paper is related to the use of corpus linguistics methods and techniques, in conjunction with the so-called local grammar formalism, to identify patterns of changes in key financial instruments. Surprisingly a small vocabulary comprising verbs and prepositions together with orthographs like percentage and speech marks, are used to pattern information about changes in the instruments, which are attributed to a person. We showed how local grammar can be used to identify patterns of change and patterns of proper nouns. The fusion of the two local grammars also provides a powerful tool for extracting attributable financial information.

We have attempted to demonstrate that instead of invoking the full paraphernalia and concomitant expectations related to universal grammars and short contrived texts, it is perhaps better to focus on local grammars on large so-called real world corpora. Our work is entering its evaluation phase currently where the systems we have developed have been evaluated by experts in the financial information field and in language engineering. These small-scale evaluation studies have been encouraging. We have instituted our own long term evaluation which involves continually adding news data and extracting information about change in the value of the financial instrument and checking manually whether the predictions are correct or not.

We believe that the local grammar patterns can be updated in the light of new and varied reporting styles. A study of such adaptive systems is currently under way. This work is an attempt to supplement the intuition used to build local grammars and to describe IE templates.

References

Humphreys, K.; Demetriou G. & Gaizauskas R. (2000). *Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures.* In Proceedings of the Pacific Symposium on Biocomputing (Hawii 2000).

Yangarber, R. & Grishman, R. (1997). *Customization of information extraction systems*. In Paola Velardi, editor, International Worlshop on Lexically Driven Information Extraction, pp 1-11, Frascati , Italy, July 1997.

Harris, Z. (1991). *A Theory of Language and Information: A Mathematical Approach*. Oxford: Clarendon Press.

Gross, M. (1993). *Local Grammars and their Representation by Finite Automata*. In (Ed) Hoey, M. Data, Description, Discourse: Papers on the English Language in Honour of John McH Sinclair. HarperCollins Publishers. pp 26-38.

Gross, M. (1997). *The Construction of Local Grammars*. In Finite-State Language Processing, E. Roche & Y. Schabès (eds.), Language, Speech, and Communication, Cambridge, Mass.: MIT Press, pp. 329-354.

Barnbrook, G. & Sinclair, J. (1995). *Parsing CoBuild Entries*. In (Ed.) Sinclair, J.; Hoelter, M. & Peters, C. The Languages of Definition: The Formalization of Dictionary Definitions for Natural Language Processing. Luxembourg: Office for Official Publications of the European Communities. pp 13-58.

Choi, Key-sun & Nam, Jee-sun. (1997). *A Local-Grammar-based Approach to Recognizing of Proper Names in Korean Texts*. In Joe Zhou & Kenneth Church (eds.), Proceedings of the Fifth Workshop on Very Large Corpora, ACL/Tsing-hua University/Hong-Kong University of Science and Technology, pp. 2730-288.

Sinclair, J. (1991). *Corpus Concordance Collocation*. Oxford University Press.

Ranchhod, E.; Mota, C. & Baptista, J. (1999). *A Computational Lexicon of Portuguese for Automatic Text Parsing*. SIGLEX'99: Standardizing Lexical Resources.

Mason, O. (2004). *Automatic Processing of Local Grammar Patterns*. In proceedings of the 7th Annual CLUK (the UK special-interest group for computational linguistics) Research Colloquium.

Sinclair, J. (1996). Collins COBUILD Grammar Patterns 1: Verbs. HarperCollins, Glasgow.

Smadja, F. (1994). *Retrieving Collocations from Text: Xtract*. In Armstrong, S. (Editor) Using Large Corpora. London: Mit Press.