LoLo: A System based on Terminology for Multilingual Extraction

Yousif Almas
Department of Computing
University of Surrey
Guildford, Surrey, GU2 7XH, UK
y.almas@surrey.ac.uk

Khurshid Ahmad
Department of Computer Science
Trinity College,
Dublin-2. IRELAND
kahmad@cs.tcd.ie

Abstract

An unsupervised learning method, based on corpus linguistics and special language terminology, is described that can extract time-varying information from text streams. The method is shown to be 'language-independent' in that its use leads to sets of regular-expressions that can be used to extract the information in typologically distinct languages like English and Arabic. The method uses the information related to the distribution of N-grams, for automatically extracting 'meaning bearing' patterns of usage in a training corpus. The analysis of an English news wire corpus (1,720,142 tokens) and Arabic news wire corpus (1,720,154 tokens) show encouraging results.

1 Introduction

One of the recent trends in (adaptive) IE has been motivated by the empirical argument that annotated corpora, either annotated automatically or annotated manually, can provide sufficient information for creating the knowledge base of an IE system (McLernon and Kushmerick, 2006). Another equally important trend is to use manually selected seed patterns to initiate learning: In turn, active-training methods use seed patterns to learn new related patterns from unannotated corpora. Many of the adaptive IE systems rely on the existing part-of-speech (POS) taggers (Debnath and Giles, 2005) and/or syntactic parsers (Stevenson and Greenwood, 2005) for analysing and annotating text corpora. The use of corpora in IE, especially adaptive IE, should, in principle, alleviate the need for manually creating the rules for information extraction.

The successful use of POS/syntactic taggers is dependent on the availability of the knowledge of (natural) language used by the authors of documents in a given corpus. There is a wealth of POS taggers and parsers available for English language, as it has been the most widely used language in computational linguistics. However, this is not the case for strategically important languages like Arabic and Chinese; to start with, in Chinese one does not have the luxury of separating word-tokens by a white space and in Arabic complex rules are required to identify morphemes compared to English. The development of segmentation programs in these languages has certainly helped (Gao et al., 2005; Habash and Rambow, 2005). More work is needed in understanding these languages such that the knowledge thus derived can be used to power taggers and parsers.

Typically, IE systems are used to analyse news wire corpora, telephone conversations, and more recently in bio-informatics. The first two systems deal with language of everyday communications –the *general language*- whereas bio-informatics deals with a specialist domain and has its own 'special language'. English special languages, for example *languages of law, commerce, finance, science & technology*, each have a limited vocabulary and idiosyncratic syntactic structures when compared with English used in an everyday context. The same is true of German, French, Russian, Chinese, Arabic or Hindi. It appears that few works, if any, take advantage of the properties of special language to build IE systems.

Our objective is to use methods and techniques of IE in the automatic analysis of specialist news that streams in such a way that information extracted at an earlier period of time may be contradicted or reinforced by information extracted at a later time. The *impact of news* on financial and commodity markets is of consider-

Proceedings of the Workshop on Information Extraction Beyond The Document, pages 56–65,
Sydney, July 2006.

able import and is often called *sentiment analysis*. The prefix 'sentiment' is used to distinguish this kind of analysis from the more quantitative analysis of assets (called fundamental analysis) and that of price movements (called technical analysis). There is a great deal of discussion in financial economics, econometrics, and in the newly emergent discipline of *investor psychology* about the impact of 'good' and 'bad' news on the behaviour of both investors and brokers. Three Nobel Prizes have been awarded on the impact of market (trader and investor) sentiment on the value of shares, currencies, derivatives and other financial instruments (Shiller, 2000). Financial news, in addition to e-mails and blogs, has contributed to the catastrophic failures of major trading institutions (Mackenzie, 2000; Hardie & Mackenzie, 2005).

One of the key proponents of *news impact analysis* is the Economics Nobel Laureate Robert Engle who has written about asymmetry of information in a market – the brokers have more knowledge than any given individual, rumours have different impact on different actors in the market. Engle's statistical analysis suggests that the 'bad' news has longer lasting effect than 'good' news (Engle, 1993). Usually, sentiment analysis is carried out using news *proxies* which include dates/times and the names of agencies releasing key items of financial data (Anderson et al., 2002) or data like the age of a firm, its number of initial public offerings, return on investment, etc. These proxies are then regressed with share, currency or commodity prices. News impact analysis is moving into its next phase where the text of news is analysed albeit to a limited extent (Cutler et al., 1989; Chan, 2003). The analysis sometimes looks at the frequency distribution of pre-specified keywords –directional metaphors like *rose/fall, up/down,* health metaphors like *anaemic/healthy* and animal metaphors like *bullish/bearish.* A system is trained to correlate and to *learn* the changes in distribution of the prescribed metaphorical keywords, together with names of organisations, to the changes in the value of financial instruments (Seo et al., 2002; Omrane et al., 2005; Koppel and Shtrimberg 2004).

We are attempting to create a language-informed framework for news impact analysis using techniques of corpus linguistics and special language analysis. The purpose is to automatically extract patterns from a corpus of domain specific texts without prescribing the metaphorical keywords and organisation names. This, we believe, can be achieved by looking at the *lexical signature* of a specialist domain and extracting collocational patterns of the individual items of the lexical signature. The lexical signature includes key vocabulary items of the domain and names of people, places and things in the domain. There are instances in the part-of-speech tagging literature (Brill, 1993) and in IE where a corpus is used and words within a grammatical category help to extract rules and patterns comprising essential information about a domain or topic (Wilks, 1998; Yangarber, 2003). Brill, Wilks and Yangarber *induce* grammars of a universal kind: we focus on inducing a local grammar that deals with the patterning of the items in the signature. Note that in all these cases of grammar induction the intuition of the grammar builder plays a critical part whether it be in the choice of syntactic transformation rules (Brill 1993), or in choosing sense taggers and implicitly semantic rules (Wilks, 1998; Ciravegna and Wilks, 2003), or in choosing user supplied seed patterns (Yangarber, 2003). Most of the work in grammar induction is focussed on English or typologically similar languages. We have deliberately chosen typologically different languages (English and Arabic) to evaluate the extent to which our method of 'grammar induction' is language independent.

We describe a method for building domain specific IE systems: the patterns used to extract domain specific information are the N-gram collocation patterns of domain specific terms. The patterns are extracted from un-annotated domain-specific text corpora. We show how one can analyse the N-gram patterns and render them as *regular expressions*.

The thesaurus used to identify domain specific words is itself constructed automatically from a (training) special-language corpus. The frequency distribution of domain specific terms in a special language corpus shows characteristic differences from the distribution of the same terms in a general language corpus. There is little or no difference in the distribution of the so-called grammatical or closed class terms in a special and a general language corpus.

Furthermore, amongst the domain specific terms, a few tend to dominate the frequency distribution – the so-called lexical signature of a domain. These signature terms are used as nucleates for compound terms in a domain. The occurrence of the signature terms, either on their own or in a compound or a phrase, is equally idiosyncratic in that these dominant single or

compound terms co-occur more frequently with one set of words than with others. The behaviour of signature terms appears to be governed by a grammar that is local to the specialism and is not elsewhere in the general language (Harris, 1991); *local grammar* is used in general language for telling times and dates in metaphorical expressions (Gross, 1997), and in the lexicography for describing the *language of definitions* of lemmas in a lexicon (Barnbrook and Sinclair, 1996; Barnbrook, 2002). The local grammar approach, rooted in the lexical signature of a given domain can be used to extract 'sentiment' bearing sentences in financial markets (Ahmad el al., 2006) or in the description of work in a scientific laboratory (Ahmad & Al-Sayed, 2005).

We introduce a system that can help in building domain specific IE systems in English and languages that are typologically distinct from English, specifically Arabic. The development of *LoLo* was inspired by Engle's pioneering work in econometrics where news impact analysis is regarded as critical to the analysis of market movement: however much of the work in financial economics relates to the correlation of the timings of news announcements rather than the content of the news stream (Ahmad et al., 2006).

LoLo can manage a corpus and extract key terms. Given the keyword list, the system then identifies collocates and selects significant collocates on well defined statistical criterion (Smadja, 1994). Finally, local grammar rules are identified and an IE system is created.

LoLo has been used to build a local grammar to extract 'sentiment' or key (changing) market events in English and in Arabic from unseen texts. The system can help visualise the distribution of extracted patterns synchronised with the movement of financial markets.

IE systems need to be adaptive, as the specialisms in particular and the world in general is changing rapidly and this change is usually reflected in language use. There is an equally important need to build cross language IE systems as information may be in different languages. The lexically-motivated approach we describe in this paper responds to the need for an adaptive, cross domain and cross language IE systems.

2 Method

For the extraction of local grammar from a corpus of special language texts it is important to focus on the keywords. The patterns in which the keywords are embedded are assumed to comprise the principal elements of a subject specific local grammar.

The manner in which we derive the local grammar is shown in the algorithm below (Figure 1).

ALGORITHM: DISCOVER LOCAL GRAMMAR

1. **SELECT** *a special language corpus (S_L, comprising $N_{special}$ words and vocabulary $V_{Special}$).*
   - i. **USE** *a frequency list of single words from a corpus of texts used in day-to-day communications (S_G comprising $N_{general}$ words and vocabulary $V_{general}$) – for example, the British National Corpus for the English language:*
     $F_{general}:=\{f(w_1),f(w_2),f(w_3)\ldots\ldots f_{Vgeneral}\}$
   - ii. **CREATE** *a frequency ordered list of words in S_L texts is computed*
     $F_{special}:=\{f(w_1), f(w_2), f(w_3)\ldots\ldots\}$
   - iii. **COMPUTE** *the differences in the distribution of the same words in the two different corpora is computed using the in S_G and S_L:*
     $Weirdness\ (w_i)= f(w_i)special/f(w_i)general* N_{general}/N_{special}$
   - iv. **CALCULATE** *z-score for the $F_{speical}$*
     $zf(w_i)=(f(w_i)-f_{av\_special})/\sigma_{special}$
2. **CREATE** *KEY a set of N_{key} keywords ordered according to the magnitude of the two z-scores*
   $KEY:=\{key_1, key_2, key_3\ldots\ldots key_{Nkey}\}$ *such that* $z(f_{keyi})$ & $z(weridness_{keyi})> 1$
   - i. **EXTRACT** *collocates of each Key in S_L over a window of **M** word neighbourhood.*
   - ii. **COMPUTE** *the strength of collocation using three measures due to Smadja (1994):*
     U-score, k, and z-score
   - iii. **EXTRACT** *sentences in the corpus that comprise highly collocating key-words ($(U,k_o,k_1)>=(10,1,1)$)* →
   - iv. **FORM** *Corpus S_L'*
   - a. *For each $Sentence_i$ in S_L':*
   - b. **COMPUTE** *the frequency of every word in $Sentence_i$.*
   - c. **REPLACE** *words with frequency less than a threshold value ($f_{threshold}$) by a place marker #;*
   - d. **FOR** *more than one contiguous place marker, use**
3. **GENERATE** *trigrams in S_L'; note frequency of each trigram together with its position in the sentences:*
   - i. **FIND** *all the longest possible contiguous trigrams across all sentences in SL' and note their frequency*
   - ii. **ORDER** *the (contiguous) trigrams according to frequency of occurrence*
   - iii. **(CONTIGUOUS) TRIGRAMS** *with frequency above a threshold form* **THE LOCAL GRAMMAR**

Figure 1. Algorithm for the acquisition of local-grammar patterns.

Briefly, given a specialist corpus (S_L), keywords are identified, and collocates of the keywords are extracted. Sentences containing key collocates are then used to construct a sub-corpus (S_L'). The sub-corpus S_L' is then analyzed and trigrams above a frequency threshold in the sub-corpus are extracted; the position of the trigrams in each of the sentences is also noted. The sub-corpus is searched again for contiguous trigrams across the sentences: The sentences are analyzed for the existence of the trigrams in the correct position – if a trigram that, for example, is noted for its frequency as a sentence initial position, is found to co-occur with another frequent trigram that exists at the next position, then the two trigrams will be deemed to form a pattern.

This process is continued until all the trigrams in the sentence are matched with the significant trigrams.

The local grammar then comprises significant contiguous trigrams that are found. These domain specific patterns, extracted from the specialist corpus S_L' (and its constituent sub-corpus) are then used to extract similar patterns and information from a test corpus to validate the patterns thus found in the training corpus. Following is a demonstration of how the algorithm works using English and Arabic texts.

2.1 Extracting Patterns in English

We present an analysis of a corpus of financial news wire texts: 1204 news report produced by Reuters UK Financial News comprising 431,850 tokens. One of the frequent words in the corpus is *percent*– 3622 occurrences, a relative frequency of 0.0084%. When the frequency of this keyword is looked up in the British National Corpus (100 million words), it was found that *percent* is 287 times more frequent in the financial corpus than in the British National Corpus – this ratio is sometimes termed *weirdness* (of special language); the weirdness of grammatical words *the* and *to* is unity as these tokens are distributed with the same (relative) frequency in Reuters Financial and the BNC. The z-score computed using the frequency of the token in the Reuters Financial is 12.64: the distribution of *percent* is 12 standard deviations above the mean of all words in the financial corpus. (The z-score computed for weirdness is positive as well). The heuristic here is this: a token is a candidate *keyword* if both its z-scores are greater than a small positive number. So *percent* -most frequent token with frequency and weirdness z-score over zero- *was* accepted as a keyword.

The collocates of the keyword *percent* were then extracted by using mutual information statistics presented by Smadja (1994). A collocate in this terminology can be anywhere in the vicinity of +/- N-words. The frequency at each neighbourhood is calculated and then used to compute the 'peaks' in the histogram formed by the neighbourhood frequencies and the strength of the collocation calculated on a similar basis. The keyword generally collocates with certain words that have frequencies higher than itself – the *upward collocates*- and collocates with certain words that have lesser frequency – the *downwards collocates* (These terms were coined by John Sinclair). Upwards collocates are usually grammatical words and downwards collocates are lexical words – nouns, adjectives- and hence the downwards collocates are treated as candidate compound words. There were 46 collocates of *percent* in our corpus – 34 *downwards* collocates and 12 *upwards* collocates. A selection of 5 downwards and upwards are shown in Table 1 and 2 respectively.

Collocate	Frequency	U-score	k-score
shares	1150	1047	3.01
rose	514	2961	2.43
year	2046	396	2.40
profit	1106	263	1.65
down	486	996	1.40

Table 1. Downward collocates of *percent* in a corpus of 431,850 words.

Collocate	Frequency	U-score	k-score
the	23157	6744	14.40
to	12190	7230	10.29
in	9768	4941	8.49
a	10657	3024	8.44
of	10123	3957	8.24

Table 2. Upward collocates of *percent* in a corpus of 431,850 words.

The financial texts comprise a large number of numerals (integers and decimals) and these we will denote as <no>. The numerals collocate strongly with *percent* for obvious reasons. The collocates are then used to extract trigrams comprising the collocates that occur at particular positions in the various sentences of our corpus:

Token A	Token B	Token C	Freq	Position
<no>	**percent**	and	16	1
rose	<no>	**percent**	18	1
<no>	**percent**	after	23	2
<no>	**percent**	of	47	2
<no>	**percent**	rise	11	2

Table 3. Trigrams of *percent*.

There are many other frequent patterns where the frequency of individual tokens is quite low but at least one member of the trigram has higher frequency: such low frequency tokens are omitted and marked by the (#) symbol. All the trigrams containing such tokens with at least two others are used to extract other significant trigrams. Sometimes more than one low frequency tokens precede or succeed high frequency tokens and they are denoted by the symbol (*) as shown in Table 4. The search for contiguous trigrams leads to larger and more complex patterns, Table 5 provides some examples.

Token A	Token B	Token C	Freq	Position
rose	<no>	**percent**	18	1
#	<no>	**percent**	29	2
#	shares	were	10	2
*	<no>	**percent**	57	2
<no>	**percent**	#	24	2

Table 4. Trigrams of *percent* with omitted low frequency words (denoted as * for multiple tokens and # for a single token).

Local Grammar Patterns	Freq
<s> the * <no> **percent**	28
<s> * rose <no> **percent**	26
<s> # shares # <no> **percent**	22
<s> * fell <no> **percent**	20
<s> * <no> **percent**	18
<s> # shares were up <no> **percent** at	17

Table 5. Some of top patterns of *percent* (<s> identifies a sentence boundary).

2.2 Extracting Patterns in Arabic

Arabic is written from right to left and its writing system does not employ capitalization. The language is highly inflected compared to English; words are generated using a root-and-pattern morphology. Prefixes and suffixes can be attached to the morphological patterns for grammatical purposes. For example, the grammatical conjunction "and" in Arabic is attached to the beginning of the following word. Words are also sensitive to the gender and number they refer to and their lexical structure change accordingly. As a result, more word types can be found in Arabic corpora compared to English of same size and type. Short vowels which are represented as marks in Arabic are also omitted from usual Arabic texts resulting in some words having same lexical structures but different semantics.

These grammatical and lexical features of Arabic cause more complexity and ambiguity, especially for NLP systems designed for thorough processing of Arabic texts compared to English. A shallow and statistical approach for IE using texts of specialism can be useful to abstract many complexities of Arabic texts.

Given a 431,563 word corpus comprising 2559 texts of Reuters Arabic Financial News and the same thresholds we used with the English corpus, *percent* (*al-meaa*, المئة) is again the most frequent term with frequency and weirdness z-score greater than zero. It has 3125 occurrences (0.0072%), a frequency z-score of 19.03 and a weirdness of 76 compared against our Modern Standard Arabic Corpus (MSAC).

There were 31 collocates of *percent*; 7 upwards and 23 downwards. The downwards collocates of *percent* appear to collocate with names of instruments i.e. *shares* and *indices* (Table 6).

The upwards collocate are with the so-called closed class words as in English like *in, on* and *that* (Table 7).

Collocate	Freq	U-score	k-score
by-a-ratio (*be-nesba*, بنسبة)	1257	39191	7.87
point (*noqta*, نقطة)	1167	9946	6.44
the-year (*al-aam*, العام)	1753	344	3.34
index (*moasher*, مؤشر)	1130	409	2.55
million (*milyoon*, مليون)	2281	600	2.32
share (*saham*, سهم)	705	206	1.84

Table 6. Downward collocates of *percent* (*al-meaa*, المئة).

Collocate	Freq	U-score	k-score
in (*fee*,في)	21236	434756	40.99
to (*ela*, الى)	3339	25145	9.81
from (*min*, من)	10344	4682	9.58
on (*ala*,على)	5275	117	3.10
that (*ann*, ان)	5130	260	2.65

Table 7. Upward collocates of *percent* (*almeaa*, المئة).

Using the same thresholds the trigrams (Table 8) appear to be different from the English trigrams in that the words of movement are not included here – this is because Arabic has a richer morphological system compared to English and Financial Arabic is not as standardised as Financial English: however, it will not be difficult to train the system to recognise the variants of *rose* and *fell* in Financial Arabic. Table 9 lists some of the patterns.

Token A	Token B	Token C	Freq	Position
<no>	in (*fee*,في)	**percent** (*al-meaa*,المئة)	197	1
in (*fee*,في)	**percent** (*al-meaa*,المئة)	*	39	1
in (*fee*,في)	**percent** (*al-meaa*,المئة)	to (*ela*, الى)	22	2
percent (*al-meaa*,المئة)	to (*ela*, الى)	<no>	21	3
#	in (*fee*,في)	**percent** (*al-meaa*,المئة)	66	4

Table 8. Trigrams of *percent* (*almeaa*, المئة).

Local Grammar Patterns	Freq
<s> * <no> في المئة * percent in	34
<s> * <no> بنسبة <no> في المئة الى <no> *by-a-ratio* *in* *percent* *to*	23
<s> # سهم # <no> في المئة *share* *in* *percent*	21
<s> # مؤشر # الاوسع نطاقا بنسبة <no> في المئة الى <no> نقطة </s> *index* *wider* *by-ratio* *in* *to percent* *point*	18
مؤشر * <no> نقطة أي <no> في المئة الى <no> نقطة </s> *index* *point* *namely* *in* *percent* *to* *point*	16
في * يوم # بنسبة <no> في المئة مع * *in* *day* *by-ratio* *in* *percent* *with*	10

Table 9. Some patterns of *percent* (*almeaa*, المئة).

3 Experimental Results

We have argued that a method that is focused on frequency at the lexical level(s) of linguistic description – single words, compounds, and N-grams- will perhaps lead to patterns that are idiosyncratic of a specialist domain without recourse to a thesaurus. There are a number of linguistic methods – that focus on syntactic and semantic level of description which might be of equal or better use.

In order to show the effectiveness of our method we apply it to sentiment analysis – an analysis that attempts to extract qualitative opinion expressed about a range of human and natural artefacts – films, cars, financial instruments for instance. Broadly speaking, sentiments in financial markets relate to the 'rise' and 'fall' of financial instruments (shares, currencies, commodities and energy prices): inextricably these sentiments relate to change in the prices of the instruments. In both English and Arabic, we have found that *percent* or equivalent is a keyword and trigrams and longer N-grams embedded with this keyword relate to metaphorical movement words– *up, down, rise, fall*. However, in English this association is further contextualised with other keywords – *shares, stocks*- and in Arabic the contextualisation is with shares and the principal commodity of many Arab states economies – *oil*. Our system 'discovered' both by following a lexical level of linguistic description.

For each of the two languages of interest to us, we have created 1.72 million token corpora. Each corpus was then divided into two (roughly) equal sized sub corpora: training corpus and testing corpus; the testing corpus is sub-divided into two testing corpora $Test_1$ and $Test_2$ (Table 10). First, we extract patterns from the Training Corpus using the *discover local grammar* algorithm (Figure 1) and also from $Test_1$. Next, the $Training_1$ and $Test_1$ corpora are merged and patterns extracted from the merged corpus. The intuition we have is that as the size of the corpus is increased the patterns extracted from a smaller sized corpus will be elaborated: some of the patterns that are idiosyncratic of the smaller sized corpus will become statistically insignificant and hence will be ignored. The conventional way of testing would have been to see how many patterns discovered in the training corpus are found in the testing corpora; we are quantifying these results currently. In the following we describe an initial test of our method after introducing *LoLo*.

Corpus	English		Arabic	
	Texts	Tokens	Texts	Tokens
$Training_1$	2408	861,492	5118	860,020
$Test_1$	1204	431,850	2559	431,563
$Training_2$ ($Training_1$+$Test_1$)	3612	1,293,342	7677	1,293,342
$Test_2$	1204	426,800	2559	428,571
Total	**4816**	**1,720,142**	**10,236**	**1,720,154**

Table 10. Training and testing corpora used in our experiments.

3.1 LoLo

*LoLo (*stands for *Local-Grammar for Learning Terminology* and means 'pearl' in Arabic*)* is developed using the .NET platform. It contains four components summarised in Table 11.

Component	Functionality
CORPUS ANALYSER	Discover domain specific extraction patterns
RULES EDITOR	Group, label and evaluate patterns and slots
INFORMATION EXTRACTOR	Extract information
INFORMATION VISUALISER	Visualise patterns over time

Table 11. Summary of *LoLo's* components.

The various components of *LoLo* –the *Analyser, Editor*, *Extractor* and the *Visualiser*, can be used to extract and present patterns; the system has utilities to change script and the direction of writing (Arabic is right-to-left and English left-to-right). Table 12 is an exemplar output from *LoLo:* "rise in profit" event patterns expressed similarly in English and Arabic financial news headlines found by the *Corpus Analyser*.

English	* profit up <no> percent
Arabic	ارتفاع أرباح * <no> في المئة *profit rise (up) percent in*

Table 12. "Rise in profit" patterns in Arabic and English where the * usually comprises names of organisations or enterprises.

The pattern acquisition algorithm presented earlier is implemented in the *Corpus Analyser* component, which is the focus of this paper. It can be used for discovering frequent patterns in corpora. The user has the option to filter smaller patterns contained in larger ones and to mine for interrupted or non-interrupted patterns. It can also distinguish between single word and multi word slots.

Before mining for patterns, a corpus preprocessor routine performs a few operations to improve the pattern discovery. It identifies any punctuation marks attached to the words and separates them. it also identifies the sentences boundaries and converts all the numerical tokens to one tag "<no>" as numbers can be part of some patterns, especially in the domain of financial news.

The *Rules Editor* is at its initial stages of development, currently it can export the extraction patterns discovered by the *Corpus Analyser* as *regular expressions*.

A time-stamped corpus can be visualised using the *Information Visualiser*. The *Visualiser* can display a time-series that shows how the extracted events emerge, repeat and fade over time in relation to other events or imported time series i.e. of financial instruments. This can be useful for analysing any relations between different events or detecting trends in one or more corpora or with other time-series.

LoLo facilitates other corpus and computational linguistics tasks as well, including generating concordances and finding collocations from texts encoded in UTF-8. This is particularly useful for Arabic and languages using the Arabic writing system like Persian and Urdu which lack such resources.

3.2 Training and Testing

3.2.1 English

We consider the English $Training_1$ corpus first. We extracted the significant collocates of all the high frequency/high weirdness words, where 'high' defined using the associated z-scores, in the training corpus. Trigrams were then extracted and high frequency trigrams were chosen and all sentences comprising the trigrams were used to form a (training) sub corpus. The sub-corpus was then analysed for extracting the local grammar.

The 10 high frequency N-grams extracted automatically from the $Training_1$ Corpus (861,492) are listed in Table 13. The $Test_1$ corpus has most of the trigrams in the $Training_1$ corpus, particularly some of the larger N-grams (Table 14).

Rank	Top 10 patterns comprising 'percent'	Freq
1	<s> the * <no> percent	45
2	<s> the * was up <no> percent at <no>, <no> </s>	33
3	<s> * <no> percent #, <no> </s>	24
4	<s> * up <no> percent	21
5	<s> the * was down <no> percent at <no> , <no> </s>	19
6	<s> * <no> percent after	18
6	<s> * <no> percent to <no> , <no> yen	18
7	<s>, # shares were up <no> percent at <no>	17
8	<s> shares in * <no> percent	15
9	<s> * rose <no> percent to <no>	14
10	<s> # shares rose <no> percent to <no>	13
10	<s> fell <no> percent to <no>	13

Table 13. Patterns of *percent* extracted from $Training_1$ corpus.

Patterns	Freq
<s> # shares # <no> percent	22
<s> shares in * <no> percent	13
<s> # shares were up <no> percent at	17

Table 14. Patterns of *percent* extracted from $Test_1$ corpus found as sub-patterns in $Training_1$.

We then merged the $Training_1$ and $Test_1$ corpora together and created $Training_2$ corpus comprising of 3612 texts and 1,293,342 tokens. The Algorithm was executed on the merged corpus and a new set of patterns were extracted, in particular the most frequent pattern in the $Training_1$ Corpus (*<s> the * <no> percent*), was elabo-

rated by the Algorithm as well as those patterns shown in Table 15.

Training$_1$ Corpus	Freq	Training$_2$ Corpus	Freq
<s> the * was down <no> percent at <no> , <no> </s>	19	<s> the * **index** was down <no> percent at <no> , <no> </s>	23
<s> the * was up <no> percent at <no>, <no> </s>	33	<s> the * **index** was up <no> percent at <no> , <no> </s>	34

Table 15. Comparison between two patterns in Training$_1$ and Training$_2$ corpora.

The patterns related to the collocations of shares and percent from Training$_1$ were preserved in Training$_2$. The test on Test$_2$ corpus showed similar results: the smaller N-grams related to the movement of instruments were similar to the Test$_1$ Corpus. The analysis of Arabic texts is shown below with similar results.

3.2.2 Arabic

Some of frequent N-grams extracted automatically from the Training$_1$ Arabic corpus (860,020) are shown in Table 16. Similar to the English corpora the Test$_1$ Arabic corpus has most of the trigrams in the Training$_1$ Corpus and some larger N-grams(Table 17).

Rank	Top 10 patterns comprising 'percent'	Freq
1	* المئة في <no> * <s> *percent in*	35
2	* في * بنسبة <no> في المئة *percent in by-ratio in*	31
3	* <no> نقطة بنسبة <no> في المئة الى <no> نقطة </s> *point to percent in by-ratio point*	28
4	<s> * في * في المئة * *in percent in*	24
4	<s> * بنسبة <no> في المئة الى <no> *to percent in by-ratio*	24
5	* في * الى <no> في المئة *percent in to in*	21
5	<s># مؤشر * نطاقا بنسبة<no>في المئة الى <no>نقطة</s> *point to percent in by-ratio zone index*	21

Table 16. Patterns of *percent* (*almeaa*, المئة) extracted from Training$_1$ Arabic corpus.

Patterns	Freq
# المئة في <no> بنسبة * *percent in by-ratio*	10
في المئة في <no> بنسبة * *in percent in by-ratio*	10
* المئة في <no> بنسبة * <s> *percent in by-ratio*	11

Table 17. Patterns of *percent* (*almeaa*, المئة) extracted from Test$_1$ Arabic corpus found as sub-patterns in Training$_1$.

After merging the Training$_1$ and Test$_1$ Arabic corpora together into a corpus of 7677 texts and 1,293,342 tokens, new set of patterns were extracted as well. Some of the frequent patterns in the training corpus were elaborated more as well like the pattern shown in Table 18 where the token *and-rise* (*wa-ertifaa*, وارتفع) was added to the pattern.

Training$_1$ Corpus	**Freq**	**Training$_2$ Corpus**	**Freq**
<s> # مؤشر # الاوسع نطاقا بنسبة <no> في المئة الى <no> نقطة </s>	13	<s> **وارتفع** مؤشر # الاوسع نطاقا بنسبة <no> في المئة الى <no> نقطة </s>	17

Table 18. Comparison between two patterns in Training$_1$ and Training$_2$ Arabic corpora.

4 Evaluation

We have used the *Rules Editor* and the *Information Extractor* to evaluate the patterns on a corpus comprising 2408 texts and 858,650 tokens created by merging Test$_1$ and Test$_2$ corpora. The Arabic evaluation corpus comprised 5118 texts and 860,134 tokens. The N-gram pattern extractor (where N > 4) showed considerable promise in that who or what went up/or down was unambiguously extracted from the English test corpus using patterns generated through the training corpus. Initial results show high precision with the longer N-grams in English (Table 19) and Arabic (Table 20).

Pattern	Precision
<ORG> shares were down <no> percent at <no>	100% (13/13)
<Movement> <no> percent to <no> , <no> yen	100% (17/17)
the **<Index>** was up <no> percent at <no> , <no>	92% (11/12)
<ORG> shares # up <no> percent at	88% (30/34)

Table 19. Patterns with high precision (English).

Pattern	Precision
مؤشر **<Index>** <no> نقطة أي <no> في المئة الى <no> نقطة *point to percent in viz point index*	100% (42/42)
مؤشر **<Index>** لاسهم # <no> نقطة # <no> في المئة *percent in point for-shares index*	100% (27/27)
<Movement> مؤشر **<Index>**الاوسع نطاقا بنسبة <no> في المئة الى<no>نقطة *point to percent in by-ratio zone wider index*	97% (33/34)
<Movement> مؤشر **<Index>** نطاقا <no> في المئة الى <no> نقطة *point to percent in zone index*	77% (27/35)

Table 20. Patterns with high precision (Arabic).

However, some patterns return many extracted information that require trimming. For example many organizations names are extracted in Arabic using the pattern shown in table 21 but they usually have the word by-a-ratio (*be-nesba*, بنسبة) attached at the end resulting in low precision.

Pattern	Precision
<ORG> rose <no> percent to <no>	36% (5/14)
في المئة <no> **<ORG>** شركة سهم **<Movement>** *percent in company share*	30% (25/83)

Table 21. Patterns with low precision in English and Arabic

Because we have used the same training thresholds for English and Arabic, the patterns in Arabic appeared without the motion words. However the system can extract these words along with the org/instrument/index names because they appear frequently as slots in the patterns.

The N-gram patterns (when $N \leq 4$) show poor results in that either such patterns found in the training corpus are not found in the test corpus, or the patterns retrieved from test corpora are at semantic variance with the same pattern in the training corpus. This suggests that there is an optimal length of individual patterns in our local grammar.

5 Afterword

The patterns extracted from the English (and Arabic) corpora confirm to an extent the view of the proponents of *local grammar,* of a special language, that there are certain words (in our case *percent, shares, index*) that appear to have a specific grammatical category in the sense that the neighbourhood of these words is occupied by a small number of other words (*up, down, fall, rise, <no>* for instance). If we were to apply the grammars typically used in part-of-speech taggers and syntactic parsing in general, the idiosyncratic behaviour of the pivotal keywords in specialist language does not become apparent: the pivotal keywords are regarded as noun phrases and the association of these phrases is with other general categories of verb phrase, adjectival phrase and adverbial phrase.

The patterns we have extracted could have been extracted with the help of a thesaurus. And, this is the question which is critical to us: how to create and maintain a thesaurus within a domain. This is illustrated in a small way by our experiment on the Training_1 corpus where the term *index* was not statistically significant for it to appear in the trigrams that populate the local grammar. However, in Training_2, the larger corpus did contain significant frequency of the term *index* for it to make into a pattern of its own. Furthermore, many of the patterns in Training_1 persisted in Training_2. Smaller N-grams persist as well in the various Training and Test corpora – these patterns in themselves act like units around which other trigrams nucleate.

The evaluation of our Algorithm is still continuing and we are in the process of setting up experiments with human volunteers, especially those with some knowledge of financial matters to evaluate the output of *LoLo*. We intend to use information retrieval metrics of recall and the various Fβ measures.

The local grammar movement has made erratic progress since its inception in the 1960's. Now, with the advent of accessible computers with substantive memories, with the advent of the Internet and the concomitant treasure of multi-lingual text deposits and text streams, one can explore the use of such grammars in addressing the major challenges in information extraction.

Reference

Ahmad, Khurshid. and Al-Sayed, Rafif. (2005) Community of Practice and the Special Language 'Ground'. In (Eds.) Clarke, S and Coakes, E. *Encyclopaedia of Knowledge Management and Community of Practice*. Hershey (PA): The Idea Group Reference.

Ahmad, Khurshid., Cheng, David. and Almas, Yousif. (2006) 'Multi-lingual Sentiment Analysis of Financial News Streams.' In *Proc. of the 1st International Conference on Grid in Finance*, Palermo.

Andersen, Torben., Bollerslev, Tim., Diebold, Francis, and Vega, Clara. (2002). 'Micro effects of macro announcements: Real Time Price Discovery in Foreign Exchange'. National Bureau of Economic Research Working Paper 8959. (Available at http://www.nber.org/papers/w8959).

Barnbrook, Geoffrey. and Sinclair, John McH. (1996) 'Parsing Cobuild Entries'. In (Eds.) John McH. Sinclair, Martin Hoelter & Carol Peters. *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.

Barnbrook, Geoffrey. (2002) *Defining Language: A local grammar of definition sentences*. Amsterdam: John Benjamins Publishers.

Omrane, Walid., Bauwens, Luc., and Giot, Pierre. (2005) 'News Announcements, Market Activity and Volatility in the Euro/Dollar Foreign Exchange Market'. *Journal of International Money and Finance*, 24 (7), pp. 1108-1125.

Brill, Eric. (1993) 'Automatic Grammar Induction and Parsing Free Text: A transformation-based approach'. In *Proc. of 31th Annual Meeting of the Association for Computational Linguistics*, Ohio.

Chan, Wesley. (2003) 'Stock Price Reaction to News and No-News. Drift and Reversal after Headlines'. *Journal of Financial Economics*, 70(2), pp. 223-260.

Ciravegna, Fabio. and Wilks, Yorick. (2003) 'Designing Adaptive Information Extraction for the Semantic Web in Amilcare'. In (Eds.) Siegfried Handschuh and Steffen Staab, *Annotation for the Semantic Web, Frontiers in Artificial Intelligence and Applications*. US: IOS Press.

Cutler, David., Poterba, James., and Summers, Lawrence. (1989) 'What Moves Stock Prices?'. *Journal of Portfolio Management*, 15(3), pp. 4-12.

Debnath, Sandip. and Giles, C. Lee. (2005) 'A Learning Based Model for Headline Extraction of News Articles to Find Explanatory Sentences for Events'. In *Proc. of the 3rd international conference on Knowledge capture*, Alberta, Canada.

Engle, Robert. and K. Ng, Victor. (1993) 'Measuring and Testing the Impact of News on Volatility', *Journal of Finance*, 48(5), pp. 1749-1777.

Gao, Jianfeng., Li, Mu., Wu, Andi. and Huang, Chang-Ning (2005) 'Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach', *Journal of Computational Linguistics*, 31(4), Cambridge, Mass.: MIT Press, pp. 531-574

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

Habash, Nizar. and Rambow, Owen. (2005) 'Arabic Tokenization, Morphological Analysis, and Part-of-Speech Tagging in One Fell Swoop'. In *Proc. of the Conference of American Association for Computational Linguistics (ACL'05)*, Ann Arbour, MI.

Halliday, Michael, A. K. (1993) 'On the language of Physical Sciences'. In (Eds.) Halliday, Michael. and Martin, J. R., *Writing Science* pp. 54-68. London: The Falmer Press.

Hardie, Iain. and MacKenzie, Donaled. (2005) 'An Economy of Calculation: Agencement and Distributed Cognition in a Hedge Fund'. (Available at http://www.sps.ed.ac.uk/staff/An%20Economy%20of%20Calculation.pdf).

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

Kittredge, Richard. and Lehrberger, John. (1982) *Sublanguage: Studies of language in restricted semantic domains*. Berlin: Walter de Gruyter.

Koppel, Moshe and Shtrimberg, Itai. (2004) 'Good News or Bad News? Let the Market Decide'. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, Palo Alto: AAAI Press, pp. 86-88.

Mackenzie, Donald. (2000). 'Fear in the Markets'. *London Review of Books,* 22(8), pp 31-32.

McLernon, Brian. and Kushmerick, Nicholas. (2006) 'Transductive Pattern Learning for Information Extraction'. In *Proc. of EACL 2006 Workshop on Adaptive Text Extraction and Mining*, Trento.

Seo, Young-Woo., Giampapa, Joseph. and Sycara, Katia. (2002) 'Text Classification for Intelligent Agent Portfolio Management'. In *Proc. of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pp. 802-803, Bologna.

Shiller, Robert. (2000) *Irrational Exuberance*. Princeton: Princeton University Press.

Sinclair, John McH. (1996) *Collins COBUILD Grammar Patterns 1: Verbs*. HarperCollins, Glasgow.

Smadja, Frank. (1994) 'Retrieving Collocations from Text: Xtract'. In (Eds.) Armstrong, S., *Using Large Corpora*. London: MIT Press.

Stevenson, Mark. and Greenwood, Mark A. (2005) 'A Semantic Approach to IE Pattern Induction'. In *Proc. of the 43rd Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 379-386, Ann Arbour, MI.

Wilks, Yorick (1998) 'Inducing Adequate Grammars from Electronic Texts', EPSRC ROPA Grant GR/K/66215 Final Report. (Available at http://nlp.shef.ac.uk/research/reports/k66215.html).

Yangarber, Roman. (2003) 'Counter-Training in Discovery of Semantic Patterns'. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pp. 343-350, Sapporo, Japan.