Event Information Extraction Using Link Grammar

Harsha V. Madhyastha
Dept. of Computer Science and Engineering
Indian Institute of Technology, Madras
Chennai, India
harsha@peacock.iitm.ernet.in

N. Balakrishnan, K. R. Ramakrishnan
Supercomputer Education and Research Center
Indian Institute of Science
Bangalore, India
{balki@serc, krr@ee}.iisc.ernet.in

Abstract

In this paper, we present a scheme for identifying instances of events and extracting information about them. The scheme can handle all events with which an action can be associated, which covers most types of events. Our system basically tries to extract semantic information from the syntactic structure given by the link grammar system [9] to any English sentence. The instances of events are identified by finding all sentences in the text where the verb, which best represents the action in the event, or one of its synonyms/hyponyms occurs as a main verb. Then, information about that instance of the event is derived using a set of rules which we have developed to identify the subject and object as well as the modifiers of all verbs and nouns in any English sentence, making use of the structure given by the link parser. The scheme was tested on the Reuters corpus and gave recall and precision even upto 100%.

1. Introduction

In recent times, the ever-burgeoning growth of information on the Internet has turned out to be an information glut rather than being a handy reference. The main reasons for this being the vast spread of the Internet and the lack of any organization of data. Under such circumstances, information filtering (IF) and extraction (IE) attain prime importance. The difference between IF and IE being the level at which they operate. While IF involves classification of documents based on the type of information they contain, IE is concerned with identifying the parts of a text related to a certain fact. The interest among research groups to build IE systems has been high since the beginning of the Message Understanding Conferences (MUCs) and Text Retrieval Conferences (TRECs) in the late 1980s. This interest was sustained by the TIPSTER program which ran through the last decade. The systems developed for these conferences performed admirably well but they were mainly based on domain-dependent rules whose formulation required painstaking effort over a long time. Recently, the onus has shifted to techniques based on wrappers and Hidden Markov Models (HMMs). Wrapper-based techniques exploit the semi-structured form of information available on the Internet. The generation of the wrapper has also been automated [3][7], making it highly suitable for use with semi-structured information. A review of wrapper-based IE schemes can be found in [6]. On the other hand, HMM-based techniques [4] operate on natural language text, making use of statistical information. Their structure can also be built without manual interference [8]. Their main disadvantage is that they require lots of training data to begin with. A survey of recent IE schemes is found in [10].

Our scheme operates on natural language text and does not require any training data. It makes use of the syntactic structure assigned to the input text by the link parser. The link grammar is a robust system which handles almost all aspects of English grammar. Although it is a dictionary-based system, it can handle sentences admirably well even if they have 1 or 2 words which are not in the dictionary and also, predict the part-of-speech for these words with a fair degree of accuracy. Surprisingly, the link grammar system has hardly been made use of for IE except in a few instances [5][2]. Even in these cases, they seem to have wrongly assumed that a subject-verb relationship is indicated only by the 'S' link (explained later in Section II). Here, we present a scheme to extract out instances of some chosen event from a set of documents. Our scheme can handle all events which are characterized by some action, which is a property of almost all events. The main component of our scheme is a set of rules which can be applied to first identify all the main verbs, *i.e.*, the verbs which truly represent the action in the verb phrase, in the text and then predict the subject for each of these. The scheme also helps to find out the object of the verb, when present, as well as the modifiers of all verbs and nouns. This would be of great use in building databases after documents have been clustered based on their theme. For example, to extract out all instances

of ‘murder’ in a set of crime-related documents, it would suffice to find all occurrences of the verb ‘kill’ or one of its synonyms/hyponyms in the text and then find their subjects, objects and their modifiers. Hence, this scheme is of great relevance in the present day world where the need to extract out information, based on a user’s query, from a seemingly infinite source of documents, is at its peak.
The rest of the paper has been divided into the following sections :

- Link Grammar System: A brief introduction of the link grammar system
- Some Important Links : Explanation of the significance of some of the links in the link grammar system
- Rules for Prediction : Rules used to identify main verbs and their subjects and objects
- Event Information Extraction : The scheme used to identify instances of events and extract information about them
- Results : Summary of the results obtained on testing the system
- Conclusion : Analysis of the results
- Suggestions : Some suggestions for future research

2. Link Grammar System

The link grammar system assigns a syntactic structure to natural language text. It is a dictionary-based system. Each word in the dictionary is associated with a set of links. A link ending with ‘+’ implies that that word has to make that link with some word to its right and similarly ‘-’ stands for a link with a word to its left. A typical entry in the dictionary is

man : D- & (O- or S+)

This means that man must make a ‘D’ link with some word to its left and make exactly one out of a ‘O’ link to its left or a ‘S’ link to its right. The dictionary also classifies the words according to their parts of speech. So, when a sentence is given as input to the link parser it searches for those words in the dictionary and tries to build a linkage structure which satisfies the following three rules:-

1. Planarity : The links do not cross when drawn above the words.
2. Connectivity : The links suffice to connect all the words of the sequence together.
3. Satisfaction : The links satisfy the linking requirements of each word in the sentence.
4. Exclusion : No two links may connect the same pair of words.

Also, the words are tagged according to their parts of speech. Nouns are tagged with ‘n’, verbs are tagged with ‘v’, prepositions are tagged with ‘p’ and so on.

3. Some Important Links

The following is a list explaining the significance of some of the important linkages of the link grammar system which have been used in our scheme:-

- A and AN : Connects pre-noun modifiers like adjectives or nouns to the following noun. eg - the **huge man** sat there, the **tax proposal** is to be revised
- B : Connects transitive verbs back to their objects in relative clauses and questions. eg - the **man** he **killed**, **what** did you **eat**. Also, connects the main noun to the finite verb in subject-type relative clauses. eg - the **teacher** who **taught** me was tall.
- DP : Connects possessive determiners to gerunds in cases where the gerund is taking its normal complement. eg - **your telling** Jane to leave was a mistake.
- I : Connects infinitive verb forms to certain words such as modal verbs and “to”. eg - he **has to** be present, they **should do** their work
- J : Connects prepositions to their objects. eg - the man **with** the **dog** is here.
- M : Connect nouns to various kinds of post-noun modifiers like prepositions and participles. eg - the **man with** the umbrella, the **lady to** whom I proposed
- MV : connects verbs and adjectives to modifying phrases that follow. eg - the man **slept in** the room, it was **hotter yesterday**
- MX : Connects nouns to post-nominal noun modifiers surrounded by commas. eg - the **man**, **who** killed him, was arrested.
- O, OD and OT : Connects transitive verbs to their objects, direct or indirect. eg - he **played cricket**, I **gave you** a book
- P : Connects forms of the verb “be” to prepositions, adjectives and participles. eg - he **is playing**, the boys **are in** the field, she **was angry**
- PP : Connects forms of “have” to past participles. eg - he **has gone**

- R : Connects nouns to relative clauses. eg - the **student who** was absent, the **dress that** she wore
- RS : Connects the relative pronoun to the verb. eg - the man **who chased** us
- S, SI, SX and SXI : Connects subject nouns to finite verbs. eg - a **child likes** sweets
- TO : Connects verbs and adjectives which take infinitival complements to the word "to". eg - they **planned to** party.

4. Rules for Prediction

At the core of our event information extraction scheme is the set of rules that we have come up with to predict the subject and object of a verb as well as modifiers of all verbs and nouns. Our subject/object prediction scheme begins once the sentence has been passed through the link parser and the linkage for that sentence has been obtained. As the link grammar requires that no two links cross each other, no two links connect the same pair of words and all the words form one unit, the linkage structure can be represented in the form of a tree. The elements of the tree are then analyzed to first find the main verbs and then if possible, find their subjects and objects.

4.1. Identifying the Main Verbs

The link parser itself tags the verbs of the sentence with a 'v' tag but all of them are not main verbs and all of them do not require subjects. Here, a main verb is considered to be the word in the verb phrase which actually represents the action done, *i.e.*, words like infinitives (eg - to, will), modal verbs (eg - must, should) and sometimes forms of "be" (like in "he was playing") are neglected. Also, verbs do not need subjects when they are acting as an adjective.

In order to identify the main verbs, all the words tagged with 'v' are considered first. Then verbs are pruned out based on the following conditions :-

1. Verbs which make an 'A' link with some noun to their right or make a 'M' link with some noun to their left without making any other link act as adjectives and so they do not need a subject. (Refer Fig. 1)

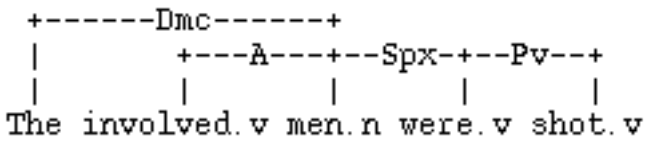

Figure 1. Verb as adjective

2. Infinitives, modal verbs and forms of "be", when followed by a verb are neglected. This is done by neglecting all words which make a 'P', 'PP' or 'I' link with some word to their right. Also, if a verb makes a 'TO' link with "to" which in turn makes a 'I' link with some word, then both are neglected. (Refer Fig. 2)

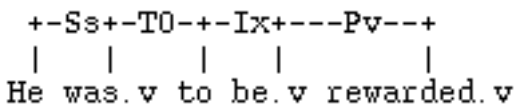

Figure 2. Pruning verb phrase

3. In some cases, adjectives are also treated as verbs because they too form 'P' links with forms of "be" and, 'MV' and 'TO' links with modifying phrases just like verbs. This is necessary to predict the subjects of verbs occurring in modifying phrases. (Refer Fig. 3)

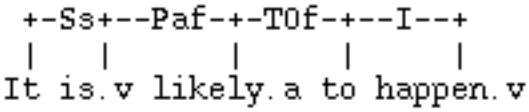

Figure 3. Adjectives as verbs

4.2. Subject and Object Prediction

After all the main verbs have been identified, the subject and object (if it exists) for each of them is predicted based on the following rules. First, lets go through the rules for subject prediction. The rules are applied in hierarchical fashion with the next rule being applied only if the subject is not found with all the rules before it. The only exception is, rule 4 is applied only if subject is found in a rule before it. Also, each rule is applied not only to the main verb identified but also to each word occurring in the verb phrase.

1. The most basic and obvious way of identifying the subject is by finding a word which makes either a 'S', 'SI', 'SX' or 'SXI' link with the verb. (Refer Fig. 4)

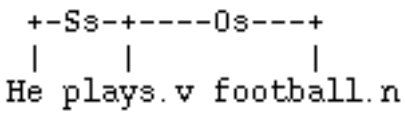

Figure 4. He → plays

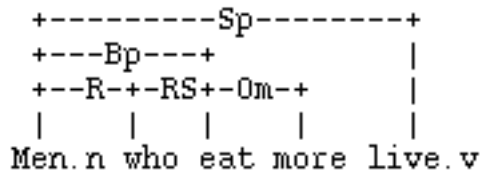

Figure 5. Men → eat

2. If a verb is connected to a noun by a 'B' link and the verb also bears a 'RS' link then the noun with which it has the 'B' link is its subject. (Refer Fig. 5)

3. The above rules do not work in the case of passive sentences as the word with the 'S' link is actually the object. A sentence is deduced as passive if a 'Pv' link is present in the verb phrase. In such sentences, the subject is usually present in the form of the phrase *"by subject"*. Or else, the object is identified as done for normal cases and classified as the subject. (Refer Fig. 6)

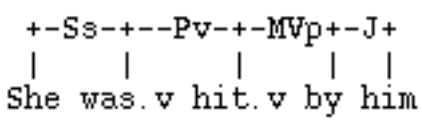

Figure 6. him → hit → She

4. In some cases, the actual subject may be connected by a 'MX*r' link to the subject found by any one of the above three rules. (Refer Fig. 7)

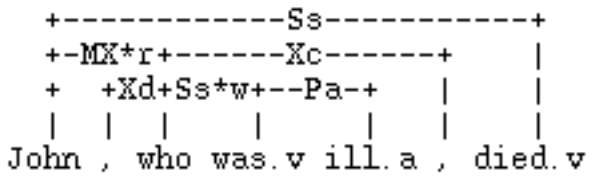

Figure 7. John → was

5. When the verb occurs in the form of a gerund, the subject may be attached to the verb with the 'DP' link. (Refer Fig. 8)

The above five rules are the basic rules for finding the subject directly.

6. If a verb is connected to the object of some other verb with 'Mg' link then that object is the subject for this verb. (Refer Fig. 9)

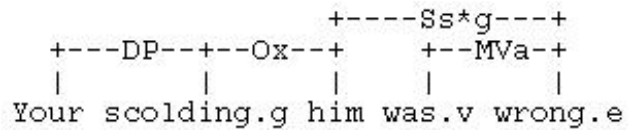

Figure 8. Your → scolding

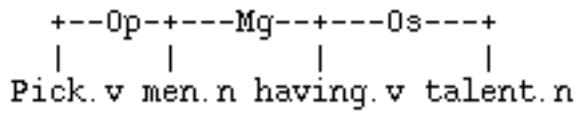

Figure 9. men → having

7. If a verb occurs in the phrase modifying a verb, wherein the phrase is connected to the verb with 'MV' link, then its subject is the subject of the verb it modifies. (Refer Fig. 10)

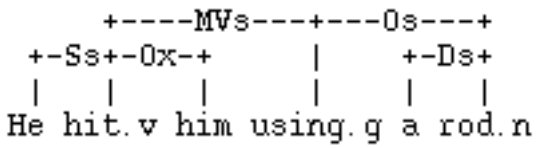

Figure 10. He → using

8. If a verb occurs in the phrase modifying a verb, wherein the phrase is connected to the verb with 'TO' link, then its subject is the object (if it exists) of the verb it modifies. If the verb which is modified does not have an object then its subject is the required subject. (Refer Fig. 11)

9. In the extreme case of all the above rules failing, the subject of the verb is taken as any noun to which the verb is connected with a 'M' link. This rule need not be correct at all times.

From the above rules it is clear that to find the subject, the object of the verb (if it exists) and the modifying phrases of both the verb and the object will also have to be found. The rules for finding the object are as follows:-

1. Here too, the most basic way of finding the object is to find the word which makes either an 'O', 'OD' or 'OT' link with the verb.

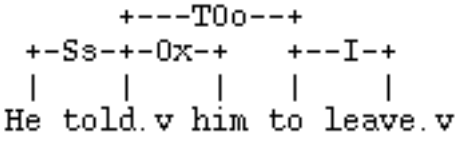

Figure 11. him → leave

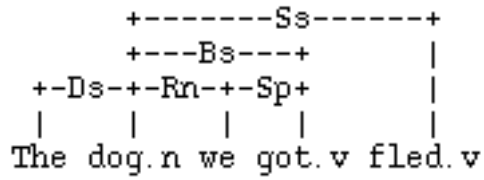

Figure 12. we → got → dog

2. If the verb makes a 'B' link with a noun and the verb does not have a 'RS' link then that noun is the object of the verb. (Refer Fig. 12)

3. If a verb makes a 'Mv' link with the object of some other verb then that object is the object of this verb as well. (Refer Fig. 13)

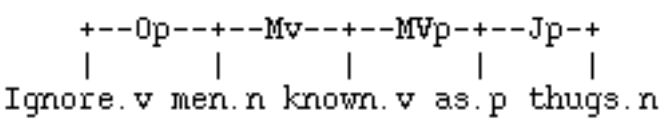

Figure 13. known → men

4. Also, as already mentioned, in the case of passive sentences, the subject and object are interchanged.

After finding the verb, subject and object, their modifiers have to be found as they are required to find the subject and object of verbs occurring later. Any phrase which forms a complete linkage structure on its own and is connected to a verb by a 'MV' or 'TO' link is classified as a verb-modifying phrase. eg - In Fig. 10, the phrase "using a rod" modifies the verb 'hit'.
Similarly, for subjects and objects, in fact for any noun, a phrase is said to modify them if it forms a complete linkage structure on its own and is connected to the noun by means of a 'M' link. eg - In Fig. 13, the phrase "known as thugs" modifies the noun 'people'.
It has to be noted that the subject may not be deducible in all cases from the information given in the article. For instance, in the sentence "*He is said to have killed him.*", it is not possible to deduce who is the subject for the verb *said* from the article alone. Such verbs are called 'agentless passives'.

5. Event Information Extraction

The inspiration behind our scheme is the modus operandi used by us, humans, to extract information from text. When we search for some event in a document, we usually first think of some key words, the presence of which we think will most probably indicate an instance of the required event. We then make a quick scan of the document, searching for the words thought of in the previous step and whenever found, we focus on that sentence and process it further. Our scheme is based on similar lines but is limited to events which can be characterized by some action. This subset in fact covers almost all kinds of events. Taking inspiration from our "instinctive" ability, our scheme follows the following steps :-

1. First, the user has to give as input some key verb which he/she thinks best represents the action which characterizes the required event.

2. Next, we take all synonyms and hyponyms of the chosen key verb. A hyponym of a word is essentially similar in meaning but is more specific.

3. Now we run the chosen documents through the link grammar parser which tags the words according to part of speech and assigns a syntactic structure to the sentence.

4. We now search for all occurrences of the verbs identified in step 2. We only select those instances where they occur as main verbs.

5. Having identified all sentences where either the key verb or one of its synonyms/hyponyms acts as a main verb, we now use the rules enumerated in the previous section to identify the subject and object (if present) of the verb as well as the modifiers of all three (verb, subject and object).

Each occurrence of the key verb, or one of its synonyms/hyponyms, as a main verb is considered to be one occurrence of the required event. So, by finding the subject, object as well as all available modifiers, almost all information about that instance of the event can be extracted from the document.

6. Results

As is pretty obvious from the scheme outlined above, the heart of the system lies in the working of the rules for prediction of subject, object and their modifiers. The rules for this scheme were derived by running the link parser on articles from various online newspapers. The newspapers were chosen from different regions ('The Times' - UK, 'Rediff' - India, 'New York Times' - USA) to account for different writing styles. Also, the articles covered different themes like weather reports, politics, statements of people and editorials. The abstracts of some papers were also used to take into consideration technical style of writing. On the whole, around 100 articles were used to ascertain that the rules did work. To test these rules on a standard set of documents,

Table 1. Results obtained for subject prediction

Topic of Article	Recall	Precision
Mergers/Acquisitions	85%	62%
Earnings/Earnings Forecasts	87%	82%
Money/Foreign Exchange	77%	50%
Money Supply	100%	77%
Trade	75%	88%

the Reuters corpus was used. In order to use the same set of articles for testing the event information extraction part as well, articles from the following 5 categories were chosen - mergers/acquisitions, earnings and earnings forecasts, money/foreign exchange, money supply, trade. From each category, 20 articles were picked at random, but making sure that each of them was atleast 50 lines long so that they would contain a reasonable amount of information.
The results for the testing of the subject prediction scheme were measured using standard information extraction units recall and precision where
Recall = (No. of verbs for which subject was identified/No. of verbs identified)
Precision = (No. of verbs for which subject was identified correctly/No. of verbs for which subject was identified)

Next, in the 100 articles chosen above (20 articles from each of the 5 categories mentioned), a search was done for all events of either “buying” or “selling” using the scheme outlined in Section IV. The verbs “buy” and “sell” were used as the key verbs for the events “buying” and “selling”, respectively. So, all synonyms and hyponyms of the verbs “buy” and “sell” were found using WordNet[1] and all occurrences of these as main verbs were determined using the link grammar structure of all sentences in each document. Considering each such occurrence as one instance of the respective event, all information about it was extracted out. The success of this scheme as well was measured using recall and precision which in this case are
Recall = (No. of instances of the event identified correctly/No. of instances of the event)
Precision = (No. of instances of the event identified correctly/No. of instances of the event identified)

An instance of the event was considered to be identified correctly if an instance of the event is indeed described in that sentence and if the subject and object (if present) were identified correctly.

Table 2. Results obtained for extraction of “buying" events

Topic of Article	Recall	Precision
Mergers/Acquisitions	100%	74%
Earnings/Earnings Forecasts	73%	60%
Money/Foreign Exchange	83%	25%
Money Supply	63%	35%
Trade	63%	60%

Table 3. Results obtained for extraction of “selling" events

Topic of Article	Recall	Precision
Mergers/Acquisitions	62%	100%
Earnings/Earnings Forecasts	81%	100%
Money/Foreign Exchange	100%	81%
Money Supply	100%	100%
Trade	70%	87%

7. Conclusion

In most articles, the cause for low recall in the subject prediction scheme was seen to be the presence of agentless passives. On the other hand, the cause for low precision was seen to be presence of verbs which have their subjects in other parts of the article rather than the sentence in which they occur. Also, to predict the subject of verbs which occur later in the sentence, the system uses the subject and object of verbs occurring before it in the sentence. Hence, if the subject or object of a verb is predicted incorrectly, the error is carried forward through the rest of the sentence.
In the case of the event information extraction testing, the need for information filtering, *i.e.*, classify documents based on their theme, is clearly shown by the low precision for “buying” events in the articles belonging to categories Foreign Exchange and Money Supply. This is because the synonyms of “buy” like “acquire” tend to denote different meanings in such situations. In the case of articles belonging to categories Acquisitions or Earnings, the verb “acquire” usually stands for “obtaining something with money”. Whereas in other domains like foreign exchange, “acquire” may stand for just “obtaining something” not necessarily with money. So, it is important that text classification be done before trying to extract information about events. But, on the whole, the recall and precision shown by both the subject prediction scheme as well as the event information extraction system are high enough to make them

feasible in real-life situations.

8. Future Work

In the subject prediction scheme, the linkage of each sentence is considered one by one. If the subject is in some other sentence as is usually the case in articles like

> Yesterday, an earthquake, of magnitude 6.0 on the Richter scale, hit the city. It is one of the worst disasters in recent times. Hundreds are feared dead and thousands more injured.

Here, the subject for injured is the earthquake but as it occurs in a different sentence, it cannot be detected by the scheme described above. To handle such instances, some kind of inter-sentence linkage structure will have to be developed.

Another area that could improve the accuracy of the system is to disambiguate co-reference (anaphora resolution), *i.e.*, find each pronoun stands for which noun in the article. For instance, in the above article, it is tough to decide "it" in the second sentence stands for earthquake unless its known that a city cannot be a disaster. This will help in finding the true subject. The subject prediction scheme could also be developed to identify 'cause and effect' relationships.

References

[1] C. Fellbaum. Wordnet, an electronic lexical database. *The MIT Press*, 1998.

[2] A. Janevski. Information extraction from university web pages. *Master's Thesis, University of Kentucky*, 2000.

[3] N. Kushmerick, D. Weld, and R. Doorenbos. Wrapper induction for information extraction. *International Joint Conference on Artificial Intelligence*, 1997.

[4] T. R. Leek. Information extraction using hidden markov models. *Master's Thesis, University of California, San Diego*, 1997.

[5] D. Moll, M. Hess, and J. Berri. An answer extraction system for unix manpages. *Technical Report, Computational Linguistics, University of Zurich*, 1998.

[6] I. Muslea. Extraction patterns for information extraction tasks : A survey. *The AAAI Workshop on Machine Learning for Information Extraction*, 1999.

[7] I. Muslea, S. Milton, and C. Knoblock. A hierarchial approach to wrapper induction. *Proceedings of the Third International Conference on Autonomous Agents*, 1999.

[8] K. Seymore, A. McCallum, and R. Rosenfeld. Learning hidden markov model structure for information extraction. *AAAI Workshop on Machine Learning for Information Extraction*, 1999.

[9] D. Sleator and D. Temperly. Parsing english with a link grammar. *Carnegie Mellon University Computer Science technical report*, CMU-CS-91-196, 1991.

[10] K. Techner. A literature survey on information extraction and text summarization. *Computational Linguistics Program*, 1997.