Automatic Template Creation for Information Extraction, an Overview

Robin Collier

CS-96-07

Abstract:

Information Extraction (IE) approaches currently assume that a template exists which sufficiently defines the requirements of the task. Substantial human effort is required to generate these basic templates and to provide a development corpus. In the two principal IE competitions, the Message Understanding Conference (MUC) and Tipster, the templates were constructed directly from the experience of analysts. This manual approach cannot always be assumed.

This proposal concerns the automatic construction of MUC-style templates, substantially reducing the human effort required. The approach will carry out a corpus-based analysis of task-relevant documents, identifying and analysing the interaction between the fundamental elements. A resource which defines semantic relationships will be necessary to identify and categorise these fundamental elements. This application is of particular interest to researchers in the field of IE and automatic abstracting.

1. Introduction

A shortcoming of current IE technology is the assumption that a template exists which sufficiently encapsulates the requirements of the task. Substantial human effort is necessary to generate these basic template structures which are task specific, and to annotate relevant documents to provide a development corpus. In the case of MUC (ARPA 1996) and Tipster (ARPA 1993), the templates were constructed directly from the experience of analysts. This manual approach cannot always be assumed as users with sufficient knowledge of the domain may not be available, or the interaction between template components may be extremely complex.

This proposal concerns the automatic construction of MUC-style templates from newspaper articles, enabling flexible template creation and substantially reducing the human effort required. The approach will carry out a corpus-based analysis of documents that are relevant to the task to identify and analyse the interaction between the fundamental elements, and create a template which sufficiently defines the application. A resource which defines semantic relationships between words will be necessary to identify and categorise these fundamental elements. This application is of particular interest to researchers in the field of IE and automatic abstracting, as it will provide a flexible and justifiable strategy for the generic development of application domains.

To consider the automatic creation of MUC-style templates it is necessary to identify the interaction that occurs between the elements of the basic template. Figure 1 provides an example of an instantiated template from the MUC-6 *management successions* domain, along with the paragraphs from the original document which describe the succession events.

Four levels of information[1] are present: the highest level concerns the *document* which contains two occurrences of relevant relationships (management successions), the second level concerns the management succession *relationships*, the third level concerns the *objects*[2] that are involved in the *relationships*, the lowest level concerns the *features*[3] which provide specific information concerning the *relationships* and *objects*.

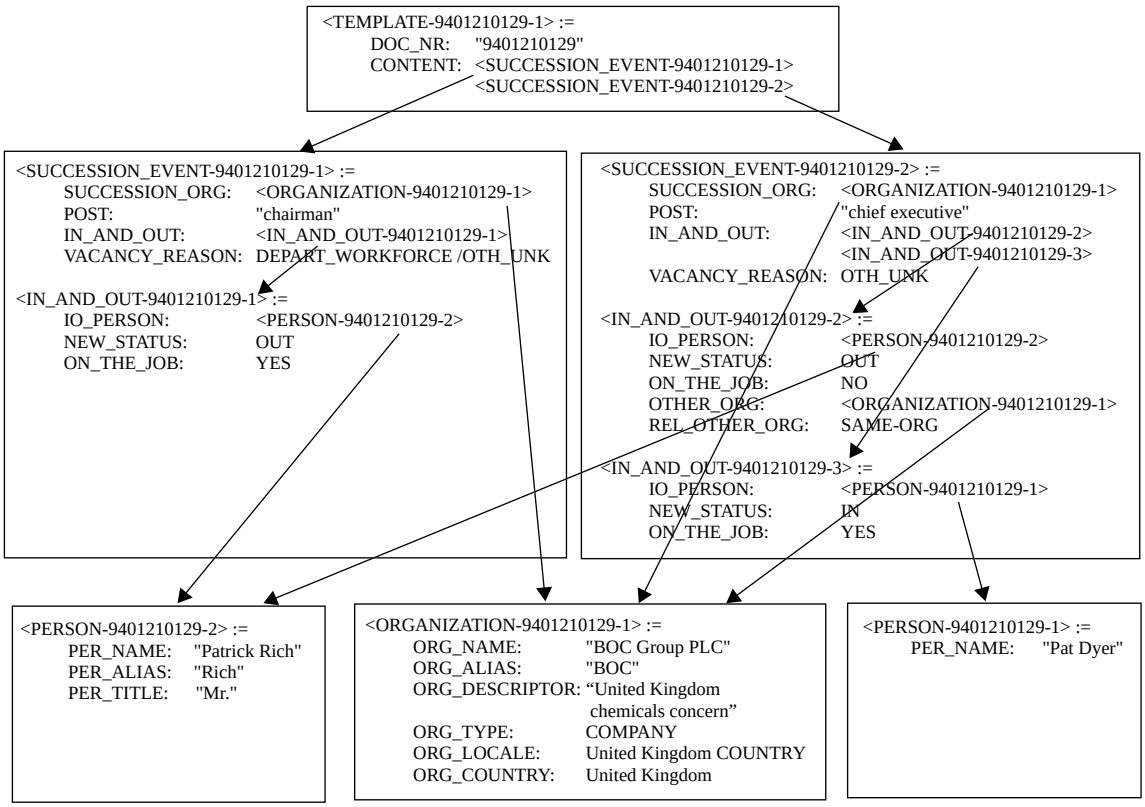
The approach discussed below will consider a template from the perspective of three fundamental types of element: *objects*, *relationships*, and *features*. The interaction between these elements will be considered in the context of the domain provided by relevant documents.

2. Background

The major IE competitions, for example, MUC (ARPA 1996), currently employ manual approaches to create templates for the particular IE domain (Hobbs & Israel 1994; Onyshkevych 1993). Even though an approach to automatically create templates would provide a significant development within IE research (for example, the Tipster architecture (Grishman et al. 1995) briefly considers *customisable extraction systems*, section A.2), relatively little research is being carried out in this area. Costantino et al. (Forthcoming) consider a semi-automatic method for template creation, a user defines the slots using a restricted form of natural language, thus enabling a formal representation of the task to be extracted. The method is at a very early stage of development, and it is unclear how automated this approach will prove to be. Mikheev and Finch (1995) identify significant concepts within a domain and generate relevant lexical items. Although this approach was not developed to generate templates for IE, there are significant similarities as it identifies the underlying features and structures within the domain that is being analysed.

In the broader domain of IE there has been substantial research carried out which has utilised manually generated templates. The LArge Scale Information Extraction system (LASIE) (Gaizauskas et al. 1996) was the Sheffield entry in the recent MUC-6 competition. The task was to identify articles

1. The word “information” refers to textual information that occurs in natural language documents, not in the broader sense used in Computer Science (for example, binary numbers/files, etc.).
2. “Objects” are referred to as entities in the Message Understanding Conferences (ARPA 1996), and are related to the MUC-6 template element subtask (this is described in the background section).
3. “Features” are referred to as slots in the Message Understanding Conferences.



BOC also said that Patrick Rich plans to step down from his post of chairman at the beginning of April, shortly after his 63rd birthday, partly because of health reasons. The board expects to make a decision on a successor “soon.”

Mr. Rich, who has worked for BOC for 11 years, had been chairman and chief executive until late last year, when the roles were split and Pat Dyer was named to the chief executive spot.

Figure 1: An instantiated MUC-6 template with relevant paragraphs of the document

from the Wall Street Journal (WSJ) that were relevant to the management succession domain and carry out four subtasks: named entity identification, coreference resolution, template element generation, and scenario template generation providing a summary of succession events occurring within each relevant article (Figure 1). The development stages of MUC-6 provided competitors with training data containing relevant documents and their associated filled templates.

The most successful approaches in the recent MUC trials have approached the problem using a pattern-matching technique by extracting common structures from training corpora which identify key constituents of the template, this is commonly referred to as pattern-matching IE. These techniques have largely been developed by American participants: the University of Massachusetts (Fisher et al. 1996), MITRE (Aberdeen et al. 1996), New York University (Grishman 1996), SRA (Krupka 1996), and SRI (Appelt et al. 1996).

Corpus-based approaches have also been applied to the MUC domain in the context of automatically generating lexicons for IE: AutoSlog[1] (Riloff 1993), CRYSTAL[2] (Soderland et al. 1995), and PALKA (Kim & Moldovan 1993).

1. AutoSlog was used by the University of Massachusetts’ MUC-4 and 5 system, CIRCUS.
2. CRYSTAL is being used by the University of Massachusetts’ current MUC system.

The AutoSlog system used an annotated training corpus of relevant documents to produce concept nodes which identify potential slot fillers in the Latin American terrorism domain (ARPA 1992). Linguistic rules were used to identify trigger words and linguistic constraints. AutoSlog-TS (Riloff & Shoen 1995) was developed to eliminate the requirement of the annotated training corpus (replacing them with documents that were relevant to the domain), and used simple statistical techniques to identify which of the prospective concept nodes were more prevalent in the relevant training texts than the irrelevant ones.

CRYSTAL used a similar approach, and was applied to the domain of hospital discharge reports, although it still relied on annotated texts. One important distinction is that CRYSTAL used a domain-specific semantic network to provide more reliable constraints on slot fillers and enable generalisation of concepts that were sufficiently close to each other in the semantic hierarchy. CRYSTAL utilised a medical MetaThesaurus and semantic network (Lindberg et al. 1993) to identify and generalise concepts in the medical domain.

Pattern-matching IE and automatic lexicon generation (in particular AutoSlog-TS) are related to the technique proposed in this paper, as the format of the template is automatically extracted by identifying regular structures and elements existing in relevant documents.

A crucial issue during template creation is the identification of semantic relationships between words, and phrases, which exist in the same context and therefore provide a prospective constituent of the template. A resource is required which is similar to the MetaThesaurus used by CRYSTAL, but is sufficiently generic to cover a wide variety of potential domains.

A widely used lexical resource is the Longman Dictionary of Contemporary English (LDOCE[1]) (Procter 1978). LDOCE contains around 60,000 entries and 36,000 word forms. The Computing Research Laboratory in New Mexico developed the Lexical Database (LDB) (Stein et al. 1993), a database populated with LDOCE entries and augmented with a noun-genus hierarchy, semantic codes, and domain codes. Four-letter domain codes indicate the categories of text in which a word sense would be expected to appear. The code can be decomposed into a pair of two-letter codes, primary and secondary. There are 300 primary codes, for example, AC-architecture and BZ-business, and 2500 secondary codes, for example, MI-military and LB-labour.

Preliminary investigation of LDB provided evidence that is was insufficient for the purposes of template creation, the detail of knowledge provided was incomplete and too generic. An example of the problems are provided by the employment senses of the words “retire” and “resign” for which the domain codes should be similar, although, the “resign” entry is blank, Table 1.

Sense Id:	resign_0_1
Domain:	__ __
Definition:	to give up (a job or position)
Sense Id:	retire_0_4
Domain:	BZ LB (business labour)
Definition:	to (cause to) stop working at one’s job, profession,...

Table 1: LDB entries for the employment senses of “resign” and “retire”

Another popular lexical resource is WordNet (Miller et al. 1990), the knowledge is hierarchically structured in the style of a thesaurus. The objective of WordNet is to provide an aid for searching dictionaries conceptually. WordNet contains approximately 95,600 different word forms (51,500 simple words and 44,100 collocations) organised into 71,100 word meanings, or sets of synonyms (synsets). The lexicon is divided into four categories: nouns, verbs, adjectives, and adverbs. Five different relationships are supported: synonymy, antonymy, hyponymy, meronymy, and entailment.

1. A detailed description of LDOCE and WordNet is provided in Stevenson et al. (1996).

Examples of the hypernyms for employment related senses of the words, “leave”, resign”, “retire”, and “step down”, are provided in Table 2. The hierarchical structure provides a method for clustering related words. Table 2 shows that ““resign”, “retire” and “step down” all intersect at the “=> leave office, quit, step down” hypernym, these also intersect with “leave” one hypernym higher, “=> leave, depart”. The structure of the hierarchies within WordNet is complex, for example, the verbs are divided into 14 lexicographer files (for example, state, event, causation), and there are a total of 573 unique verb beginners (for example, *change* in Table 2). There is significant variation in depth of these hierarchies, therefore, a similarity measure in terms of number of hypernyms between two words would require normalising by the total depth of the hierarchy.

Sense 9
leave, depart
    => change

Sense 1
vacate, renounce, **resign**, give up
    => leave office, quit, step down
        => leave, depart
            => change

Sense 1
retire, go into retirement
    => leave office, quit, step down
        => leave, depart
            => change

Sense 1
leave office, quit, **step down**
    => leave, depart
        => change

Table 2: WordNet hypernym chains

This resource identifies useful relationships between words for this application, in addition, a naive form of sense disambiguation is automatically carried out by identifying the node in the hierarchy where majority of the words intersect, illustrated by the description of Table 2.

One further resource was investigated, the Longman Activator (Summers 1993), which provided relationships similar to those in WordNet. This resource is, as yet, unused in NLP research. The Activator is a language production dictionary containing information regarding ideas and how to express them in English. The Activator contains a large number of concepts, which themselves contain subconcepts representing more specific usages of related words in a particular domain. There are no further levels of the hierarchy above these three (concepts, subconcepts and word groups). An index of words provides pointers from each word to a set of relevant subconcepts, these subconcepts are similar to the different senses of the word. An example of a concept and subconcepts is provided below to clarify the structure of the knowledge contained in the Activator.

The concept for “leave a job or organization” contains seven subconcepts, the first three being:

1. to leave a job or organization
2. to make someone leave their job
3. to be forced to leave your job

The first subconcept contains the related words: “leave”, “quit”, “resign”, “retire”, “hand in your notice/resignation”, and “pack it in/jack it in”. The second subconcept contains the words “fire”, “sack/ give sb[1] the sack”, “dismiss”, etc. Not all subconcepts contain synonymous words, some simply contain words used in the same context, for example, “president”, “chairman”, and “executive”.

Investigation of this resource suggested that it would provide similar relationships to those in WordNet. In addition, it is a dictionary and therefore supplies richer information, such as natural language definitions of the concepts (WordNet provides some, although, these are often cryptic), and examples of usage. The word indexing mechanism, and shallow hierarchy, provide a more tractable method (compared to identifying the hypernym intersections in the WordNet hierarchies, and standardising the similarity measure) for extracting relationships that occur between groups of words. The Activator is a generic lexicon, therefore, it is not tied to any particular domain, so it should provide a broad range of knowledge covering a variety of applications.

3. Overview of approach

Information Retrieval (IR) techniques were used to identify sets of task relevant documents which were distributed to participants in the MUC competitions. For this reason, it will be assumed that effective IR technology is available that will identify relevant documents, although, this assumption raises a further issue. A very general IR query, such as "I am interested in European financial news", will retrieve a set of articles that are too general for a template to be automatically generated, there simply isn't sufficient conformity between the documents for IE to be viable. Classification of the set of domains for which the technique can be applied is necessary and will result in a list of guidelines/ restrictions defining how specific the IR query should be. A clear definition of the scope of the technique is an essential prerequisite of the entire process. The corpus of relevant documents provided by the IR will provide the input to the template creation approach described below.

The description of the MUC-6 template structures (section 1) resulted in considering template creation from the perspective of identifying three fundamental elements in the context of the task:

- the *objects* which interact
- *relationships* representing the interaction between the objects
- the *features* that are specific to the objects/relationships

Template creation requires the identification of significant relationships that occur between the objects that are specific to the task. The approach will carry out corpus-based analysis to provide evidence of the three elements listed above. It will be necessary to utilise the Longman Activator to identify and classify these semantic relationships. WordNet will also be used as a secondary resource to provide semantic relationships which support and add to those found in the Activator.

A fully automatic approach to create templates is the most desirable. This would only use relevant and irrelevant documents, and would require a sufficiently detailed semantic knowledge source. If the semantic knowledge provided by the Longman Activator and WordNet is insufficient, then interaction with the user may be necessary. The primary objective is to minimise the level of interaction necessary to produce the most automated approach possible.

3.1 Object identification

The initial stage is to identify objects that are fundamental in the relevant documents. For the system to be generic to any domain, a technique is necessary to automatically identify the relevant objects in the context of the corpus.

The most difficult scenario is to identify/classify all types of entities automatically. In MUC competitions this task is referred to as *named entity identification*[1]. To identify the multitude of different entity classes occurring in the vast range of alternative domains it would be necessary to utilise enormous quantities of world knowledge, providing substantial collections of related words.

In the context of the MUC competitions this subtask has proven to be the most tractable, approaches have obtained performances of over 90% when identifying organisations, person names,

1. *sb* represents the subject of a sentence.

1. Note that some entities are not referred to by names, for example, aircraft components are referred to by part numbers in the MUC-6 aircraft orders development corpus.

currencies, locations, times and dates. The LASIE system (Gaizauskas et al. 1996), which participated in MUC-6, was one such system that obtained high performance levels when identifying named entities. The LASIE named entity identifier makes extensive use of gazetteers and word lists (obtained from language repositories), and trigger words (such as Co., Inc., and Mr.) which are incorporated into a noun phrase grammar specifically designed for named entity recognition. The success of these approaches has resulted in commercially available software (SRA 1995) which provide extremely high levels of competence for the identification of the entities occurring in electronic documents.

The availability of these resources makes the automatic identification of entities within the MUC-6 domain a tractable task. It is unnecessary to rework approaches that have already been resolved with such high performance, although it will obviously be necessary to extend the coverage of such classifiers for entities in alternative domains[1]. The only further consideration will be to identify which of the entities identified are fundamental objects within the task, and which others are less important. The less important entities will often represent features (MUC slots) associated with either an object or relationship (section 3.3), for example, the irregular occurrence of a date may provide evidence that it isn't a fundamental object, but it may relate to a feature representing the date on which a management succession occurred.

There are some well founded assumptions with which the distinction between objects and less relevant entities can be made. Due to all of the relevant documents providing at least one instance of the relationship, the fundamental objects should reside in **every** relevant text. In addition, Sheffield's experience obtained from involvement in the MUC-6 task provided evidence that those occurrences are generally related to the principal verbs which define the relationships (these are identified in the next stage). Whereas, the less important entities (and potential features) will be more sporadic, due to the shorter texts providing insufficient detail to identify all of the feature values. MUC-6 involvement also provided evidence that the features are more often associated with other (less relevant) verbs.

Discrimination between objects and potential features can be considered from two perspectives:

1. Entities which occur in every relevant document are considered to be the objects
2. A final decision on which entities are objects is withheld until further evidence is provided in the following stage which identifies the principal verbs (relationships)

3.2 Relationship restriction

The second stage of the creation of templates is to restrict the relationships to those that are specifically relevant to the context of the application. A plausible approach is to consider verb/subject/object interaction within sentences. The relationships and objects will be present in all relevant documents. If sufficient generalisation of verbs into higher order categories can be carried out, then the frequency of occurrence of these higher order categories will provide evidence concerning the significant relationships that occur within the relevant documents.

Analysis of the sentence level verb/subject/object relationships would provide a list of sentence level interactions that occur, for example:

<PERSON> retire <ORGANISATION>
<PERSON> resign <ORGANISATION>
<ORGANISATION> employ <PERSON>
<ORGANISATION> turnover <CURRENCY>

One subtask of the MUC competition is coreference resolution. Identification of coreferences between a person's name and a personal pronoun, or a company name and a pronoun, will enable replacement of the anaphors with the original antecedent. This will increase the number of explicit objects occurring in the document, and the quantity of verb/subject/object relationships identifiable.

1. This will require identification of word lists, trigger words, and the production of a noun phrase grammar for new classes of entities.

LaSIE's grammar will enable identification of verb/subject/object relationships, although, as is the case with all grammars, these will not be 100% accurate or provide full coverage. A more naive, but straightforward (and therefore computationally efficient), approach would be to identify the verb that is closest to the occurrence of an object. It is hoped that the errors introduced by the simplicity of the approach will be compensated by the quantity of results obtained by using the corpus-based approach.

These interactions, along with their frequency of occurrence, will be conflated by merging synonymous verbs. Verbs which reside in the same Activator subconcept will provide evidence of relationship classes, and a method for automatically naming the category of relationship that has been identified. An example of this conflation would be the occurrence of the first two relationships given above, *retire* and *resign*, into the Activator class *LEAVE A JOB OR ORGANIZATION 1: "to leave a job or organisation"*. If insufficient relationships hold at the subconcept level, then verb groups which occur at concept level will also provide significant relationships, although these will not be as specific. Further evidence of relationships between words can be identified by locating WordNet hierarchy nodes where the hypernym chains intersect, as was illustrated in Table 2. Equivalent WordNet-Activator concepts will be identified by comparing the sets of words associated with WordNet hypernym nodes with those for Activator subconcepts. The higher order relationships will be listed in frequency order providing empirical evidence concerning the interaction between entities and verbs.

Verification of the significance of the relationships in the context of the specific domain could be carried out by comparing the relationships that exist in relevant documents with those that exist in irrelevant ones. Relationships which appear with a high frequency in the relevant documents, and a low frequency in the irrelevant documents, will provide evidence that the relationship is both specific to the task and not generic to all tasks.

At this stage there will be two fundamental considerations: (1) whether the Longman Activator and WordNet semantic relationships are sufficient for the task, and (2) whether the nearest verb approach is adequate or a full parse is necessary. The former is the more critical. The preliminary investigation of the Activator and WordNet for the management succession domain has provided evidence that extremely useful relationships exist, although it is not expected that these resources will be sufficient for all prospective domains. Therefore, it may be necessary to calculate a confidence measure to decide whether it is necessary to consult with a user to confirm whether the relationships identified are correct.

3.3 Relevant feature extraction

It is assumed that by this stage the task will have been sufficiently defined, i.e. the objects have been identified and the relationships between them have been constrained. The final stage of the template creation requires an intensive corpus-based approach to identify the features that are relevant to each object/relationship. Broader analysis of the interaction occurring between objects in the context of relationships would enable identification of the features that provide the additional information necessary to sufficiently define the template, for example:

on **date** <PERSON> retired from <ORGANISATION>	(e.g. 12th April 93)
<PERSON> resigned as <ORGANISATION> **job_title**	(e.g. chairman)

The previous stages have considered interaction between objects and relationships occurring at sentence level. Carrying out the broader analysis necessary to identify the features would require consideration of paragraph level interaction. In addition, a larger corpus of relevant documents would be necessary due to the quantity and variety of features provided by a document being dependent on the level of detail that it expresses, whereas, the objects and relationships were inherently present in all of the relevant texts. Extraction of further relevant WSJ documents will be necessary from the Penn Treebank (Marcus et al. 1993).

Before considering statistical analysis of the corpus to identify potential slots, the previous stages should be reconsidered as they may provide some useful results. The entities that were identified in section 3.1, but were not classified as fundamental objects, may represent features associated with objects or relationships, for example, the *date* on which a management succession event occurred.

Coreference chains from the previous stage will include other references as well as pronouns and possessives, for example, coreference chains referring to a person may also contain references to them as the “chairman”, “chief executive”, “director”, etc. Querying the Activator with these words/phrases would identify that they are related in subconcept *MANAGER 2: “one of the top managers who control and direct the work of a company or other organisation”*.

The previous stage may have identified semantic relationships between verbs in more than one Activator subconcept, for example, *LEAVE A JOB OR ORGANIZATION 1: “to leave a job or organisation”* and *JOB/WORK 10: “to give someone a job*”. Identifying the distinction between these two concepts would generate a feature that is equivalent to the MUC-6 management succession slot *NEW_STATUS: IN* and *NEW_STATUS: OUT.*

Identification of further features using a corpus-based approach will consider the occurrence of n-grams within the relevant documents associated with syntactic information provided by the Brill tagger (Brill 1994), for example, the frequent occurrence of “post of NNP” will provide further evidence that a category exists defining the job title. Three common approaches are used to identify collocational information: analysing the raw text[1] (e.g. “Pat Dyer was named”), the part-of-speech tags (e.g. “NNP NNP VBD VBN”), or the two combined (e.g. “Pat NNP Dyer NNP was VBD named VBN”). This research will also consider other combinations, for example, replacing nouns with their part-of-speech tags and leaving other words as they occur in the raw text (e.g. “NNP NNP was named”).

The Virtual Corpus (VC) approach (Nagao&Mori 1994) will be used to identify the words/phrases that occur with a high frequency in relevant documents. N-grams which also occur frequently in irrelevant documents will not be task-specific, therefore, further analysis will be necessary to verify whether they are irrelevant. This approach could be augmented using the algorithm developed by Collier (1994) in which the relative positions of n-grams are stored. This would enable the ranking of collocations due to their distance from the objects/relationships in the paragraph.

A further example of how syntactic collocations may identify features, is that multiple occurrences of phrases, such as “promoted to chairman” and “promoted to director”, will provide evidence that the verb *promoted* significantly collocates with the preposition *to*. This indicates that a prospective category follows the preposition. Gerald Gazdar (University of Sussex) and Robert Gaizauskas (University of Sheffield) have analysed WSJ articles from the Penn Treebank (Marcus et al. 1993), deriving collocational information of this type (this work is currently unpublished).

Statistically motivated approaches such as the likelihood ratio test (Dunning 1993), mutual information (Church et al 1991), and the χ^2 test (Hoel 1971; Fienberg 1977), could also be used to identify domain-specific terms and collocations occurring within the relevant documents. This will enable concentration of the technique on terms for which there is a higher probability of relevance.

One class of feature slots used in the MUC competitions that is not identified by the approach suggested above is *set fill: to be filled by selection from a predefined list of categories*, for example, “BOC Group PLC” is classified as a *company,* rather than *government* or *other.* Classification information is particularly useful when querying databases[2], however, the categorisation of information such as company names is a difficult task. LaSIE uses predefined lists of known company/ government names, and sets of company/government trigger words. This solution is not possible during template creation, as the task is to identify the classification, therefore, it cannot by definition be predefined. Neither the Longman Activator nor WordNet contain proper nouns, although, it is feasible that trigger words such as “Ministry of” may be derivable from them.

The level of user interaction necessary at this final stage is dependent on the sufficiency of the approach and lexical resource, as well as the level of detail required. With a sufficiently detailed lexical resource, it should be possible to automatically identify majority of the relevant features, although the detection of set fill features may be beyond it’s scope and require human intervention.

---

1. Preprocessing strategies, such as replacing words with their morphological root and changing upper-case letters to lower-case, have not been included so that the example remains simple.
2. Database population is one of the primary applications of IE.

4. Evaluation and applications

It is proposed that the approach will be developed using the MUC-6 evaluation corpus in the domain of management successions. The portability of the approach will be evaluated by considering the terrorism events domain, as used in MUC-4 (ARPA 1992). The issue concerning how to evaluate the success of the automated approach is a complex one, a recent review of evaluation within NLP is provided by Sparck Jones and Galliers (1996). It is proposed that metrics that are well established in IR and IE will be used: recall, precision, the and F-measure (Chinchor & Dungca 1996; ARPA 1996a). Evaluation of automatically created templates by comparing with those developed manually (in competitions such as MUC) will be a more tractable task than evaluating templates in domains which IE hasn't been applied to, although, complexities will still arise when scoring slots that are partially correct, or others that were identified automatically but were not present in the manual template.

Evaluation of the system over a broader spectrum of domains is dependent on the sufficiency of three resources: the named entity identifier, the Longman Activator, and WordNet. Increasing the coverage of the named entity identifier to locate objects in earlier MUC competitions will be a tractable task as there is a significant overlap between entities in different competitions. This would enable evaluation of the system in alternative MUC domains. Porting the system to a completely different domain would require significant development to obtain word lists, identify trigger words, and develop the noun phrase grammar to incorporate the novel categories of objects.

The effectiveness of the Longman Activator will become apparent during the development of this application. It is assumed that the Activator will provide sufficient information to produce a substantially automatic technique for creating templates. The additional support provided by WordNet should enhance this approach. The Activator is a domain independent resource, although, the level of detail is unlikely to be sufficient for all potential domains. It will be interesting to carry out the technique on an alternative domain to evaluate the Activator's generic capabilities.

One side-effect of this approach to template creation, is that it would provide substantial support for the automatic annotation of templates (the generation of training data). The current approach used by MUC is entirely manual and is extremely labour intensive. The approach outlined above analyses the interaction between objects, relationships and features, and provides techniques which identify categories of potential slot fillers. The category members could be used to analyse relevant texts and provide an annotator with suggested template instantiations, indicating the positions in the text from where the slot fillers were extracted.

A further application of the research would be to consider the architecture defined in the Tipster documentation (Grishman et al. 1995). The section concerning *automatic template customisation* provides a limited definition of the task and requires further development. This project could help to clarify some of the problems involved in defining a generic approach and develop more detailed definitions of the objects and processes which interact during the automatic creation of templates.

This report has proposed an automatic approach for the creation of templates for Information Extraction. The approach has been implemented up to the relationship restriction stage. The relevant feature identification stage is currently being implemented and will be completed within the next few months.

5. Acknowledgments

Thank you to my supervisor, Yorick Wilks, for discussions concerning development of the template creation idea and approach, and to Kevin Humphreys for technical advice concerning LaSIE and opinions on the methodology. Also, to Yorick Wilks, Robert Gaizauskas, and Peter Croll, for feedback on earlier drafts of this report.

6. References

Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P. & Vilain, M. 1996. Description of the ALEMBIC system used for MUC-6. Proceedings of the *Sixth Message Understanding Conference* (MUC-6), pp. 141 to 155. San Francisco, California: Morgan Kaufmann.

Appelt, A., Hobbs, J.R., Bear, J., Israel, D., Kameyama, M., Kehler, A., Martin, D., Myers, K. & Tyson, M. 1996. SRI International FASTUS system MUC-6 test results and analysis. Proceedings of the *Sixth Message Understanding Conference* (MUC-6), pp. 237 to 248. San Francisco, California: Morgan Kaufmann.

ARPA. 1992. Advanced Research Projects Agency, Proceedings of the *Fourth Message Understanding Conference (MUC-4)*. San Mateo, California: Morgan Kaufmann.

ARPA. 1993. Advanced Research Projects Agency, Proceedings of the *Tipster Text Program Phase 1*. San Mateo, California: Morgan Kaufmann.

ARPA. 1996. Advanced Research Projects Agency, Proceedings of the *Sixth Message Understanding Conference (MUC-6)*. San Francisco, California: Morgan Kaufmann.

ARPA. 1996a. Appendix B: MUC-6 test scores. Proceedings of the *Sixth Message Understanding Conference* (MUC-6), pp. 291 to 316. San Francisco, California: Morgan Kaufmann.

Brill, E. 1994. Some advances in rule-based part of speech tagging. Proceedings of the *Twelfth National Conference on Artificial Intelligence* (AAAI-94). Cambridge, Massachusetts: AAAI/MIT Press.

Chinchor, N. & Dungca, G. 1996. Four scorers and seven years ago. Proceedings of the *Sixth Message Understanding Conference* (MUC-6), pp. 33 to 38. San Francisco, California: Morgan Kaufmann.

Church, K., Gale, W., Hanks, P. & Hindle, D. 1991. Using statistics in lexical analysis. In Zernik, U. (Ed.), *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Collier, R. 1994. N-gram cluster identification during empirical knowledge representation generation. Proceedings of the *Fifteenth International Conference on Computational Linguistics (COLING-92)*, pp. 1054 to 1058.

Costantino, M., Morgan, R.G. & Collingham, R.J. Forthcoming. Financial information extraction using pre-defined and user-definable templates in the LOLITA system. To appear in the *Journal of Computing and Information Technology* (CIT).

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Journal of the Association for Computational Linguistics*, vol. 19, no. 1, pp. 61 to 74.

Fienberg, S.E. 1977. *The Analysis of Cross-classified Categorical Data*. Cambridge, Massachusetts.: MIT Press.

Fisher, D., Soderland, S., McCarthy, J., Feng, F. & Lehnert, W. 1996. Description of the UMass system as used for MUC-6. Proceedings of the *Sixth Message Understanding Conference* (MUC-6), pp. 127 to 140. San Francisco, California: Morgan Kaufmann.

Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H., & Wilks, Y. 1996. University of Sheffield: description of the LaSIE system as used for MUC-6. Proceedings of the *Sixth Message Understanding Conference* (MUC-6), pp. 207 to 220. San Francisco, California: Morgan Kaufmann.

Grishman, R. 1996. The NYU system for MUC-6 or where's the syntax? Proceedings of the *Sixth Message Understanding Conference* (MUC-6), pp. 167 to 175. San Francisco, California: Morgan Kaufmann.

Grishman, R., Dunning, T, & Callan, J. 1995. Tipster Phase II architecture design document (Tinman architecture) Version 1.52. Available at *http://www.cs.nyu.edu/tipster*.

Hobbs, J. & Israel, D. 1994. Principles of template design. Proceedings of the *Human Language Technology Workshop*. San Francisco, California: Morgan Kaufmann.

Hoel, P.G. 1971. *Introduction to Mathematical Statistics* (Fourth Edition). New York: Wiley.

Kim, J. & Moldovan, D. 1993. Acquisition of semantic patterns for Information Extraction from corpora. Proceedings of the *Ninth IEEE Conference on Artificial Intelligence for Applications*, pp. 171-176. Los Alamitos, California: IEEE Computer Society Press.

Krupka, G. 1996. Description of the SRA system as used for MUC-6. Proceedings of the *Sixth Message Understanding Conference* (MUC-6), pp. 221 to 235. San Francisco, California: Morgan Kaufmann.

Lindberg, D., Humphreys, B., & McCray, A. 1993. Unified medical language systems. *Methods of Information in Medicine*, vol. 32, no. 4, pp. 281-291.

Marcus, M.P., Santorini, B. & Marcinkiewicz, R. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, vol. 19, no. 2, pp. 313 to 330.

Mikheev, A. & Finch, S. 1995. Towards a workbench for acquisition of domain knowledge from natural language. Proceedings of *European Chapter of the Association for Computational Linguistics* (EACL-95).

Miller, G.A., Beckwith, R., Fellbaum, C, Gross, D. & Miller, K. 1990. WordNet: an on-line lexical database. *International Journal of Lexicography*, vol. 3, no. 4 pp. 235 to 312.

Nagao, M. & Mori, S. 1994. A new method of n-gram statistics for large number n and automatic extraction of words and phrases from large text data of Japanese. Proceedings of the *Fifteenth International Conference on Computational Linguistics (COLING-92)*, pp. 611 to 615.

Onyshkevych, B. 1993. Template design for Information Extraction. Proceedings of the *Fifth Message Understanding Conference* (MUC-5), pp. 19 to 23. San Francisco, California: Morgan Kaufmann.

Procter, P. (Ed.) 1978. *Longman Dictionary of Contemporary English*. Harlow, England: Longman Group.

Riloff, E. 1993. Automatically constructing a dictionary for information extraction tasks. Proceedings of the *Eleventh National Conference on Artificial Intelligence* (AAAI-93), pp. 811 to 816. Cambridge, Massachusetts: AAAI Press/MIT Press.

Riloff, E & Shoen, J. 1995. Automatically acquiring conceptual patterns without an annotated corpus. Proceedings of the *Third Workshop on Very Large Corpora*, pp. 148-161.

Soderland, S., Fisher, D., Aseltine, J., & Lehnert, W. 1995. CRYSTAL: Inducing a conceptual dictionary. Proceedings of the *Fourteenth International Joint Conference on Artificial Intelligence* (IJCAI-95), pp. 1314-1321. San Mateo, California: Morgan Kaufmann.

Sparck Jones, K. & Galliers, J.R. 1996. *Evaluating Natural Language Processing Systems.* Lecture notes in Artificial Intelligence, 1083. Berlin, Germany: Springer-Verlag.

SRA. 1995. SRA International unveils "NAMETAG". *SRA Corporation Press Release.*

Stein, G.C., Lin, F., Bruce, R., Weng, F. & Guthrie, L. 1993. The development of an application independent lexicon: LexBase. *Technical Report MCCS-92-247*, Computing Research Laboratory, New Mexico State University.

Stevenson, M., Azzam, S., Catizone, R. & Collier, R. 1996. An introduction to lexical resources at Sheffield: LDOCE and WordNet. *Internal Technical Report*, Natural Language Processing group, Department of Computer Science, University of Sheffield, England.

Summers, D. (Ed.) 1993. *Longman Language Activator.* Harlow, England: Longman Group.