Introduction

The extraction of information from unstructured/semi-structured text is an important, though difficult, problem in natural language processing. Here, the extraction of information refers to the extraction of entities and the relations between them. Many different techniques exist for information extraction. This project involved the implementation of one such technique: a kernel based SVM system, due to Zelenko et al. This system requires access to a shallow parse of the text in question; consequently, an important part of this project was the implementation of two shallow parsing systems, one that uses region algebra and another that uses a CRF-based NER system.

An overview of the system is shown in the figure below.

Shallow Parse Tree

The Kernel-based SVM algorithm, as described by the Zelenko paper, works with shallow parse trees of text. It requires, as inputs, for instance, nodes to contain POS information such as whether a word is a Verb, an Adj, etc. It also requires sentences to be represented hierarchically. The sentence “John Smith is the chief scientist of the Hardcom Corporation”, for example, may be parsed into 3 nodes "John Smith", "is", "scientist", where the node "scientist" is another tree that contains the information "the chief scientist of the Hardcom Corporation". Aside from the POS tagging, trees must also contain entity tags such as "Person", "Organization" and "Location".
To parse sentences into parse tree, we decided to create a simple regex parser (chunker). We could arguably, instead of using a chunker, parse sentences into full arbitrary-depth parse trees using, for instance, a Context-Free Grammar parser. However, it is, first of all, difficult to create good CFG grammars. It is also relatively more inefficient to use a parser, since the chunker is configured to have a fixed depth. While the chunker does not guarantee a grammatically correct representation, a shallow partial parse should be sufficient for the learning algorithm.

```python
# default grammar

NP: {...}
PRP: {...}
VP: {...}
S: {...}

def getChunker(grammar = g):
    return nltk.RegexpParser(grammar, loop=5)
```

The grammar mainly tries to group things into Noun Phrases, Prepositional Phrases and Verb Phrases. For instance, the phrase "(All', ABN'), ('the', 'AT'), ('cakes', 'NN')" would be a Noun Phrase, etc. A few sample parses are listed below:

```python
parseSentence('Life is good')
parseSentence('NLP is a very good class')
parseSentence('Peak \'s Server is very stable')
parseSentence('go away')
parseSentence('He sometimes goes to the cafe')
parseSentence('All the cakes have gone')
parseSentence('Professor Berwick gave us an extension')
parseSentence('Extensions are good for you')
parseSentence('Final exams are too much fun')
parseSentence('John is a scientist of the Hardcom Corporation')
```
trees generated by the chunker (each word on the flattened tree should be in order with the corresponding word in the raw text). The tree that the matching tokens belongs to will be rewritten using the entity types as recognized by the NER system.

For instance, the sentence "John is a scientist of the Hardcom Corporation" will be chunked into:

(S
  John/NP
  (VP
    is/BEZ
    (NP a/AT scientist/NN)
    (PRP of/IN (NP the/AT Hardcom/NN Corporation/NN-TL)))))

The Stanford NER system recognizes "John" and "Hardcom Corporation" to be a "Person" and "Organization", respectively, and we then rewrite the tree into the following:

(S
  John/PERS
  (VP
    is/BEZ
    (NP a/AT scientist/NN)
    (PRP of/IN (ORGANIZATION the/AT Hardcom/NN Corporation/NN-TL)))))

The tree is then passed into the SVM classifier for relation extraction.
Stanford NER

The last step before passing inputs into the Kernel-based SVM classifier involves modifying the tags for named-entities. To do that, one implementation uses the CRF-based Stanford NER system.

The Stanford Named-Entity Recognition system is a unique and sophisticated system that uses Conditional Random Field to classify named entities. The probabilistic model uses a set of feature functions, such as word windows, surrounding POS tags, word shapes, etc. By using a CRF model, the Stanford system is able to treat features conjugationally (as opposed to how Hidden Markov Models assumes features to be independent). Another feature that makes the Stanford NER system special is how it is able to incorporate non-local information into named-entity classification (using Gibbs sampling). For instance, the Stanford system tries to keep label consistent throughout.

Our implementation is set up such that, after running the NER system on the raw data, the recognized entities are matched with the shallow
is/BEZ
   (NP a/AT scientist/NN)
      (PRP of/IN (NP the/AT Hardcom/NN Corporation/NN-TL)))
>>> special_print(s)
S
BEGIN
NODE John NP
VP
BEGIN
NODE is BEZ
NP
BEGIN
NODE a AT
NODE scientist NN
END
PRP
BEGIN
NODE of IN
NP
BEGIN
NODE the AT
NODE Hardcom NN
NODE Corporation NN-TL
END
END
END
END
END
Email: twillia5@enron.com - 503-464-3730

Enron Wholesale Services - Office of the Chairman

From: Mark Prevert, Chairman & CEO
Mark Haedicke, Managing Director & General Counsel

Subject: Confidential Information and Securities Trading

To keep pace with the fluid and fast-changing demands of our equity trading activities, Enron Wholesale Services ("EWS") has recently revised its official Policies and Procedures Regarding Confidential Information and Securities Trading ("Policies and Procedures"). These revisions reflect two major developments: (1) our equity trading activities have been extended into the United Kingdom, and (2) in an effort to streamline the information flow process, the "Review Team" will play a more centralized role, so that the role of the "Resource Group" is no longer necessary. You are required to become familiar with, and to comply with, the Policies and Procedures. The newly revised Policies and Procedures are available for your review on LegalOnline, the new intranet website maintained by the Enron Wholesale Services Legal Department. Please click on the attached link to access LegalOnline: http://legalonline.corp.enron.com/chinesewall.asp

If you have already certified compliance with the Policies and Procedures during the 2001 calendar year, you need not re-certify at this time, although you are still required to review and become familiar with the revised Policies and Procedures. If you have not certified compliance with the Policies and Procedures during the 2001 calendar year, then you must do so within two weeks of your receipt of this message. The LegalOnline site will allow you to quickly and conveniently certify your compliance on-line with your SAP Personnel ID number. If you have any questions concerning the Policies or Procedures, please call Rob Bruce at extension 5-7780 or Donna Lowry at extension 3-1939.

Enron Wholesale Services - Office of the Chairman

[From/IN1]: Mark Prevert, Chairman & CEO
Mark Haedicke, Managing Director & General Counsel

Subject: Confidential Information and Securities Trading

To keep pace with the fluid and fast-changing demands of our equity trading activities, Enron Wholesale Services ("EWS") has recently revised its official Policies and Procedures Regarding Confidential Information and Securities Trading ("Policies and Procedures"). These revisions reflect two major developments: (1) our equity trading activities have been extended into the United Kingdom, and (2) in an effort to streamline the information flow process, the "Review Team" will play a more centralized role, so that the role of the "Resource Group" is no longer necessary. You are required to become familiar with, and to comply with, the Policies and Procedures. The newly revised Policies and Procedures are available for your review on LegalOnline, the new intranet website maintained by the Enron Wholesale Services Legal Department. Please click on the attached link to access LegalOnline: http://legalonline.corp.enron.com/chinesewall.asp

If you have already certified compliance with the Policies and Procedures during the 2001 calendar year, you need not re-certify at this time, although you are still required to review and become familiar with the revised Policies and Procedures. If you have not certified compliance with the Policies and Procedures during the 2001 calendar year, then you must do so within two weeks of your receipt of this message. The LegalOnline site will allow you to quickly and conveniently certify your compliance on-line with your SAP Personnel ID number. If you have any questions concerning the Policies or Procedures, please call Rob Bruce at extension 5-7780 or Donna Lowry at extension 3-1939.