Treating Information Update Tasks with Word Clusters and Ranking Models

By Pham Quang Nhat Minh

A thesis submitted to

School of Information Science,

Japan Advanced Institute of Science and Technology,

in partial fulfillment of the requirements

for the degree of

Master of Information Science

Graduate Program in Information Science

Written under the direction of

Professor Akira Shimazu

March, 2010

Treating Information Update Tasks with Word Clusters and Ranking Models

By Pham Quang Nhat Minh (0810054)

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Professor Akira Shimazu

and approved by
Professor Akira Shimazu
Associate Professor Kiyoaki Shirai
Professor Hiroyuki Iida

February, 2010 (Submitted)

Treating Information Update Tasks with Word Clusters and Ranking Models

by

Pham Quang Nhat Minh (0810054)

School of Information Science

Japan Advanced Institute of Science and Technology

February 09, 2010

Keywords: Information Update, IR, Word Clustering, Ranking Models

Abstract

The task of updating information is a significant task in the context that many applications require documents to be updated quite often. In legal domain, it is an important task because of the massive number of legal updates and the cross-reference problem. Our research copes with a special case of the information update task, the information insertion task which aims to determine the most appropriate location to insert a piece of new information into an existing document.

In [6], the information insertion task was formulated as a hierarchical ranking problem. Each document is represented as a hierarchy of sections, paragraphs. Then, the insertion is operated over that hierarchical tree. To determine the best paragraph in the document to add a new sentence, all paragraphs of the document are ranked by a ranking function computed for each insertion sentence/paragraph pair and then, the paragraph with the highest score will be chosen. The ranking function for each insertion sentence/paragraph pair is computed based on a weight vector learned from training data. The training procedure was implemented in an online learning framework with the Perceptron algorithm [13, 8].

We investigated ranking models for the information insertion task on two datasets: Wikipedia insertion dataset obtained from [6] and Legal dataset built by ourselves. The Legal dataset was built from the United States Code which is a compilation and codifi-

cation of general and permanent federal law of the United States. The results show that when the deep semantics analysis for texts is not performed, the ranking models with the supervised approach outperform the unsupervised methods for the information insertion task.

In Natural Language Processing, semantic relations between words can be exploited when measuring semantic text similarity of two text segments. In our research, we proposed a method of measuring topical overlap between two text segments, which incorporates word clusters [5, 21, 24], and used these similarity measures as additional semantic features in the learning model. In our method, first, word clusters are derived from unlabeled data. Then, extracted word clusters are used as intermediate representations of words to exploit the semantic similarity and semantic relatedness between words which are different in surface forms but semantically related. The semantic text similarity scores are computed with various kinds of similarity functions. Our results show that combining cluster-based features with baseline features can boost the performance of the information insertion task on two datasets. In the best setting, we obtained 40.4% accuracy of choosing paragraphs on Wikipedia dataset and 52.3% accuracy of choosing section on Legal dataset.

Acknowledgement

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. In the first place I would like to express my sincere gratitude to my supervisor, Professor Akira Shimazu for his supervision, advice, assistance and guidance during the whole period of my Master's course. He has provided me kind encouragements and supports not only in my research and but also in my life.

I would like to acknowledge the Ministry of Education, Culture, Sports, Science and Technology - Japan (MEXT) for their finance support in my study time in Japan.

I have furthermore to thank Assistant Professor Nguyen Le Minh for his advice, support and reviews. During my research, Mr. Nguyen Le Minh has worked with me as a second supervisor. I have learned very much from his research experience and problem solving methodology.

I would like to thank Associate Professor Kiyoaki Shirai, Assistant Professor Makoto Nakamura, and all members of Shimazu and Shirai Laboratory in JAIST for their useful comments and valuable hints which helped me a lot to enhance the quality of my research.

I sincerely acknowledge Ms. Nguyen Thi Phuong Ha and Mr. Nguyen Thanh Son for their careful editing in English style and grammar.

Last but not least, I would like to give my special thanks to my family members, especially my parents whose encouragements and sharing enabled me to complete this work. Without my family's encouragement, I would not have finished the degree.

Bibliographic Notes

Portitions of this thesis are based on the following papers:

[1] Minh Quang Nhat Pham, Minh Le Nguyen, and Akira Shimazu. (2009). *Incremental Text Structuring with Word Clusters*. In Proceedings of the Conference of the Pacific Association for Computational Linguistics 2009, Hokkaido, Japan, pp. 109-114.

[2] Pham, Minh Quang Nhat., Nguyen, Minh Le., Shimazu, Akira. (2010). *The Information Insertion Task with Intermediate Word Representation*. The 16$^{\text{th}}$ NLP Annual Meeting, Tokyo, 2010, March.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nowadays, we are living in the society in which information is endlessly updated. For example, editors of newspapers always have to revise articles or write new ones when new information becomes available; personal websites are modified as status of individuals changes; legal documents need updating regularly, etc. With a rapid growth of World Wide Web and shared community resources on the internet, the task of updating information efficiently for very large text databases as new information emerges is a challenging problem. In the English version of Wikipedia, in average, there are more than four million edits per month in 2008[1]. The task of updating Wikipedia articles requires much time and human efforts. Therefore, some tools that aid collaborative updating or automatically perform updates could decrease maintenance efforts and potentially improve the document quality.

In legal domain, updating information is an important task. Since legal documents stipulate rules of social behaviors, they need to be revised regularly to respond changes in the society or in organizations. However, the task of updating legal documents requires much time and human effort because of two reasons. First, the number of legal documents that we have to deal with when updating legal documents is very large. For example, large organizations whose daily activities are based on a set of legal documents, always have to process a massive number of legal documents and numerous legal updates. Secondly, one

[1]http://stats.wikimedia.org/EN/TablesWikipediaEN.htm

legal document often has relations with others or refers to related legal documents. Such kind of cross-reference raises a problem when updating legal documents that one revision in a document may lead to change requirements in related documents.

The demands of supporting systems to assist legislators in making and updating law documents may increase because of some trends. First, the availability of the increasing number of online electronic law databases allows users to easily access law documents. Secondly, nowadays, e-government is becoming more and more popular. E-government or digital government is a term used to refer to the use of information and communication technology to provide and improve services, transactions and interactions with citizens, businesses, and other arms of governments [40]. With e-government, the interactions between citizens and the government have become much easier than before. Finally, the set of technologies called Web 2.0 is making e-government more practical, because both Web 2.0 and e-government are about to build communities and connect people. We believe that with the development of Web technologies, citizens who are end users of information systems can play more important roles in building the policies of governments or organizations in the near future.

In recent years, a new research field called Legal Engineering was proposed in order to achieve a trustworthy electronic society [17, 18]. One of the issues in Legal Engineering is to study methods of examining and verifying whether a law is updated consistently for its revisions by translating legal sentences into logical forms [20, 28]. This task remains challenging and it is beyond the scope of our thesis. Nevertheless, it demonstrates the significance of studying the information update task in legal domain.

With the above overview, we can see that the information update task is a very challenging problem. In our master thesis, we deal with the information insertion task which is a special case of the information update task. The information insertion task was proposed by Chen et al. [6]. More specifically, the information insertion task is to add a piece of new information into an existing hierarchically structured document while guaranteeing that new information is topically close to the surrounding context of the insertion location, and the continuity and coherence of the original document are preserved.

Chen et al. [6] modeled the information insertion task as a hierarchical structure ranking problem in which all candidate locations will be ranked by a ranking function when determining the best location to insert the new information. The location with the highest score will be chosen.

In the information insertion task, semantic text similarity measures between new information and the surrounding context of a location point were used as topical overlap features in the supervised learning framework. Semantic text similarity is an important concept in Natural Language Processing and has been applied in many tasks. For example, the text similarity score has been used to rank documents given a query in Information Retrieval [1] or to identify the central sentence of each cluster of sentences in centroid-based text summarization [35], etc.

In [6], only surface representations of words were used when measuring the topical overlap between new information and surrounding context of candidate insertion locations. However, that method cannot exploit relations between words which are semantically related. Therefore, it is attractive to consider using the intermediate representations of words rather than the words themselves to exploit the semantic similarity between words.

The idea of using intermediate representations of words has been applied successfully in many NLP tasks, such as Named-Entity Recognition tasks [27, 24], Chinese Word Segmentation [24] and recently, Dependency Parsing [21]. In these researches, first, word clusters were extracted from unlabeled data and then, semantic features based on word clusters were incorporated into supervised learning models to improve the performance of the tasks.

In this thesis, we study methods of using intermediate word representations based on word clusters to capture topical closeness between two text segments, and then apply these methods to the information update task. Our approach is as follows. First, we extract features based on intermediate representations of words, and then incorporate these features into the learning model.

We conducted experiments on two data sets: Wikipedia insertion dataset obtained from [6] and the Legal dataset built by ourselves. In present, there is no available Legal

dataset for the information insertion task. Thus, we proposed a method of building the insertion dataset from raw texts in the United States Code data [42], a law database of the United States.

In short, the contributions of our thesis are as follows:

1. We built the Legal dataset for the information insertion task, a special case of the updating task.

2. We investigated the effect of existing processing models for the Legal dataset. Due to the differences between Legal data and Wikipedia insertion data, processing models and features were modified to adapt to the new data.

3. We extended the method of using features based on word clustering for the information update task. Our experiment results on two datasets showed improvements of the method against baselines.

Our thesis is organized in six chapters. In Chapter 1, we introduce the motivation, purpose of our research, and main contributions. Chapter 2 gives the description of the information update task. In Chapter 3, we formally present the existing processing methods for the information update task with some discussions. Our proposed method is presented in Chapter 4. We describe experiments and results in Chapter 5, and finally, we give the conclusions in Chapter 6.

Chapter 2

The Information Update Task

In this chapter, we will describe the information update task, and discuss challenges of the task. Next, we will present datasets, and give a brief description of the method proposed in [6]. Finally, we will present some related works with our research.

2.1 The Task

2.1.1 Task Description

When editing a document given a piece of new information, we often find in the document sections related to the information, and then update found sections while preserving the coherence and the continuity of the text. In general, updating operations include adding new information into the document, deleting or modifying existing information in the document. The content of the document after updating must be consistent with given new information. In our viewpoint, with the current status of natural language processing techniques, building a fully automatic system for updating a document, whose quality is comparable with human editors, may not become a reality in the near future.

In this thesis, we study on a special case of the updating task, the information insertion task which was introduced by Chen et al., [6]. The research of Chen addressed the task of inserting a piece of new information into an existing document while preserving the

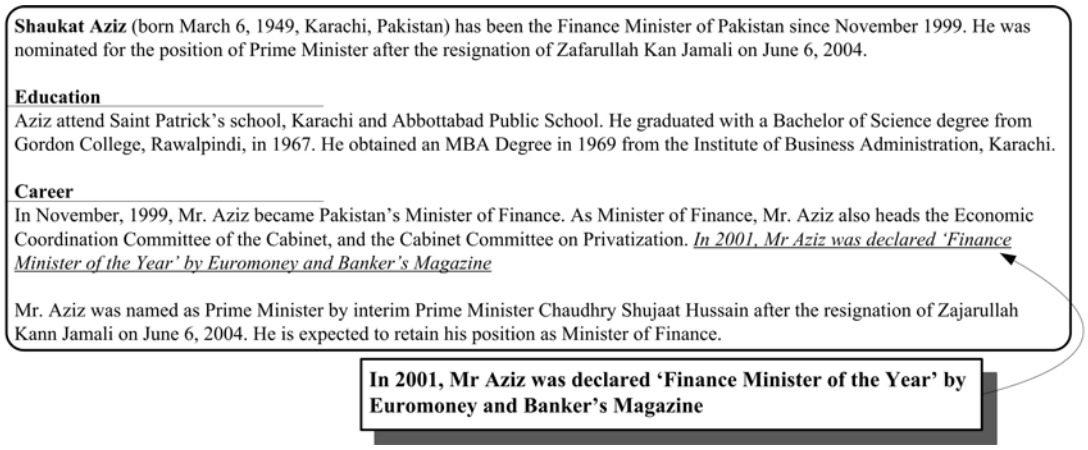


Figure 2.1: An example of Wikipedia insertion [6]

continuity and the coherence of the original text. The sentence insertion model was proposed for the information insertion task with the assumption that new information is presented in a sentence. More specifically, given a hierarchically structured document composed of sections and paragraphs, the purpose of the sentence insertion model is to determine the best paragraph to place the new sentence. Experiments were conducted on the dataset built from Wikipedia articles. Figure 2.1 shows an example of Wikipedia insertion.

The main challenge of the information insertion task is to preserve the continuity and the coherence of the original text after inserting the new sentence. In [6] these properties were maintained by examining sentences adjacent to each potential insertion point, and by directly modeling hierarchical structures of documents in the processing model.

With the above task setting, one may naturally come up with issues about the reality of the insertion task. First, when editing a document, we normally read the document carefully before modifying it, so we produce sentences and insert them into the document after we determine insertion points for them. Second, who is supposed to produce the new information? These issues seem contrast with the task setting in which we are given a new sentence and document, and we need to determine its insertion location in the document.

We argue the reality of the information insertion task. First, essentially, the task is to insert new information into an existing document. There are many ways to represent information (e.g., set of words, logic expressions, etc), and a sentence is one of the ways. Second, we interpret the meaning of the word *insertion* in a less strict meaning. Insertion means identifying the appropriate location in a document to place new information presented in a sentence. The surrounding context of the found location must be related to the new information. After the insertion location for the new sentence was determined, the sentences and surrounding sentences of the insertion location may be changed depending on editors' purpose. Third, studying the information insertion task is a useful and significant research in the case of collaboratively edited text database systems in which a document is edited by many editors, and documents are quite long. We can imagine that in such a system, new information is collected from various parties and the system will assist editors to update documents in the database. Finally, experiments in [6] were conducted on the data derived from Wikipedia articles about biography, which belong to the category "Living People". This data is a kind of special data, because new information is normally statuses of a certain individual (e.g., an event related to her/him), which change drastically over time.

In short, the information insertion task was formulated under following assumptions:

- Hierarchical structures of documents are known in advance.
- New information is presented in a sentence, a common way to convey information.
- Only one insertion location is retrieved.

What is the text coherence?

Coherence is a property of well-written texts that makes them more readable than a sequence of randomly organized sentences [22]. Normally, some documents of which components seem to "hang together" better than others are said to be coherent and therefore easier to read [43]. We need to distinguish two concepts: text coherence and text cohesion. Although cohesive devices such as lexical repletion or word usage can contribute

to the text coherence, they are different concepts. Text cohesion relates the components' interconnectedness of a text at surface level while text coherence relates to relations of the components underline the surface text, namely cognitive relations. For example, two sentences may have the cause and result relation; state or event asserted by the first sentence causes the state or event asserted by the second sentence [16]. Therefore, capturing the text coherence requires the deep semantic analysis of the input texts.

There are two levels of the text coherence: local coherence and global coherence. The local coherence captures the text coherence at the sentence level or the transition from sentence-to-sentence while the global coherence captures the text coherence of a document as a whole. Obviously, the local coherence is necessary to the global coherence.

As discussed above, the constraint of the text coherence makes the information update task challenging. In fact not many coherence factors at the deep semantic level were captured in the task. In [6], the text cohesion and the chronological order of information were exploited.

2.1.2 The Information Update Task in Legal Domain

As discussed so far, it is indispensable to study the information update task in legal domain due to the importance of legal documents and the massive amount of legal updates. In the areas of legislation, generally, there are two ways to amend statutes: enlargement and consolidation [30]. The former is to add a new provision into the existing statutes, and the latter is to revise the statutes word by word according other provisions enacted as amendment statutes. We consider the former way of updating in our thesis.

For the information insertion task in legal domain, we relax the constraint of text coherent and text continuity. The task is to find the section for new information to add with the constraint that the potential section is topically close to the new information. Legal documents normally are very long with many sections, so this relaxation makes the information update task in legal document more practical in the real world.

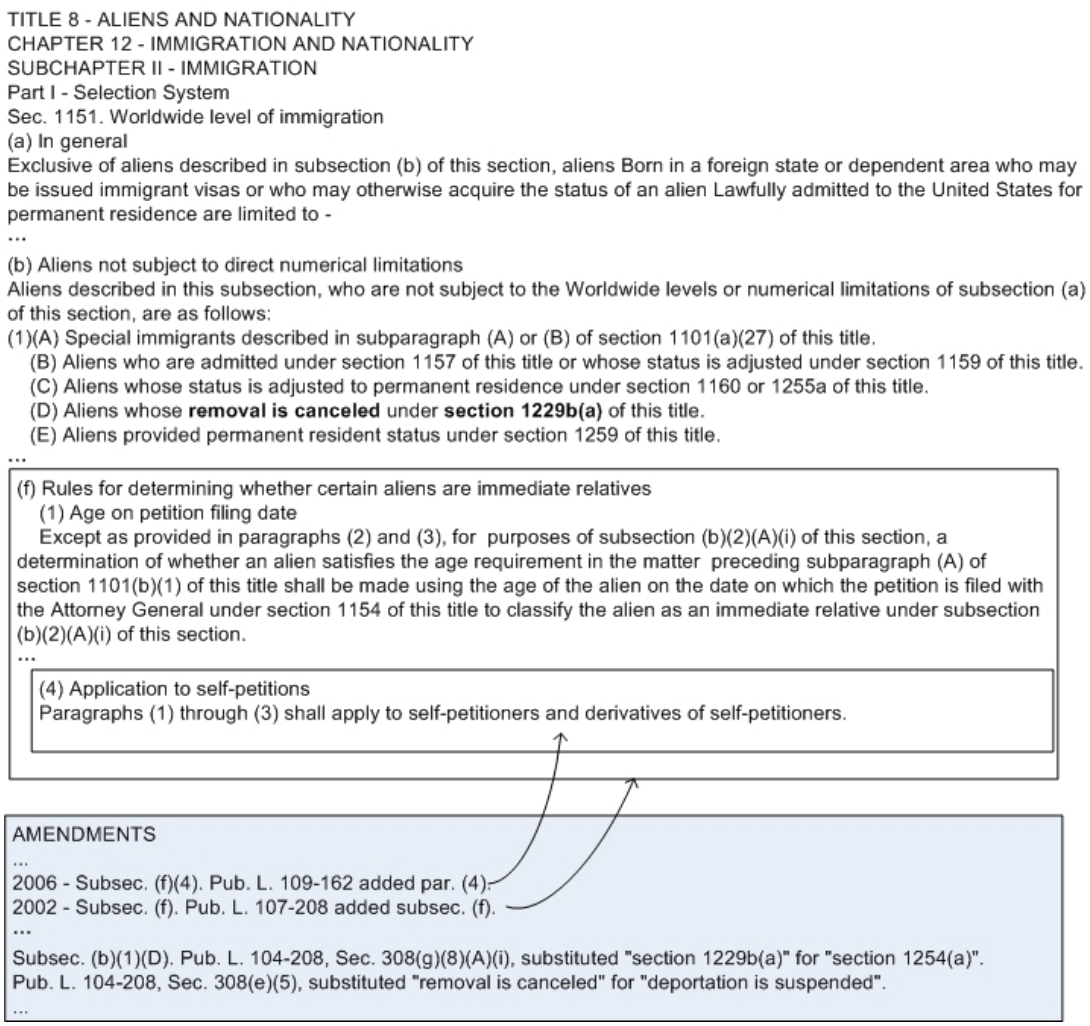


Figure 2.2: Example of two types of amendements in the U.S Code (amendements are recorded in the amendment parts after each section). Enlargement: Adding subsections (f). Consolidation: Substituting words and phrases in the document. Substituted words and phrases are in bold-face font.

2.2 Datasets

Experiments in our research were conducted on two datasets: the Wikipedia dataset obtained from [6], and the Legal dataset built by ourselves for the update task in legal domain.

2.2.1 Wikipedia Dataset

In [6], the data of insertion were obtained from the update history logs of Wikipedia articles of the category "Living People". The log records an article before and after each change in the article. From this information, the location of every inserted sentence can be identified. Totally, the Wikipedia dataset consists of 4051 insertion/article pairs from 1503 Wikipedia articles.

2.2.2 Legal Dataset

In our best understanding, currently, there is no dataset for the information update task in legal domain, so we built Legal dataset by ourselves. The Legal dataset was built from the United States Code data[1]. The details of building Legal dataset for information update task will be presented in Chapter 5 of this thesis.

2.3 Existing Methods

In [6], the information insertion task was formulated as a hierarchical ranking problem. Each document is represented as a hierarchy of sections, paragraphs, and the insertion is operated over that hierarchical tree. Features are extracted for each layer of the hierarchy. To determine the best paragraph in the document to add a new sentence, all paragraphs of the document are ranked by a ranking function computed for each insertion sentence/paragraph pair and then, the paragraph with the highest score will be chosen. The score for each insertion sentence/paragraph pair is computed based on a weight vector

[1]The plain text version is available on http://uscode.house.gov/lawrevisioncounsel.shtml

learned from training data. In [6], the training procedure was implemented in an online learning framework with the Perceptron algorithm [13, 8].

2.4 Related Works

2.4.1 Measuring Text Similarity

There are many Natural Language Processing tasks applying text similarity. For instance, in Information Retrieval [1], documents are ranked in descending order by their relevance score to an input query, and the relevance score is generally computed based on text similarity of each document to the query. Text similarity was used to identify the central sentence of each cluster of sentences in centroid-based text summarization [35], or recognizing textual entailment in RTE task [3, 11, 14].

The typical approach of computing text similarity between two text segments is to use the simple lexical matching method, namely producing the text similarity score based on overlap level of lexical units in two input segments. In order to improve this simple method, some term weighting schemes have been proposed such as or BM25 weighting scheme [36], TF-IDF weighting scheme [1], etc. The drawback of lexical matching-based methods is that they fail to identify semantic text similarity of two text segments in which lexical matching does not appear [9].

Many works have been conducted to overcome drawbacks of lexical matching based methods. Latent Semantic Analysis (LSA) method [12] which aims to find related terms in large text collections, applies Singular Value Decomposition to transform term-document matrix approximately, and then uses the transformed matrix for measuring text similarity. Corley and Mihalcea [9] attempted to combine semantic word-to-word similarity metrics into text-to-text metric, and reported the improvement against baseline lexical matching methods when applying the new text similarity metric for the task of recognizing paraphrase and textual entailment.

In our research, we propose the method of measure text similarity using intermediate representation of words in two input text segments, and investigate effects of this method

to the information update task. We expect that our proposed method can exploit semantic relationship of words which are different in their surface forms.

2.4.2 Information Retrieval

Text retrieval task (IR task) [1] aims to find documents in a large text collection, which are relevant to users' information needs given by an input query. Documents in the text collection are ranked based on a ranking function which measures relevance score between a document and the query, and documents with high score will be returned. There are some approaches to Information Retrieval. Vector Space Model [1] performs term-weighting on the query and documents, represents them in vector forms, and then computes cosine similarity score of each document with the query based on their vector representations. The Language Modeling approach [33] builds the language model for each document and ranks documents in descending order of the likelihood to generate the input query from a document.

The main problem with the word-based Vector Space Models is that relevant documents which do not contain any of the query terms cannot be retrieved. This problem suggests that queries need to be reformulated to retrieve relevant information. Some approaches were proposed for formulating queries. User Relevance Feedback [1] approaches were based on feedback information from users to reformulate queries through query expansion or term reweighting. Other approaches [34, 15, 25] attempt to perform query expansion or indexing based on a similarity thesaurus.

In [15], Gonzalo et.al showed that using WordNet synsets as indexing space instead of word forms can improve text retrieval. The constraint is that the disambiguation technique used in the disambiguation step for choosing the correct synset for each term in both documents and the input query performs well enough. Mihalcea and Moldovan in [25] proposed a semantic indexing technique that combines the benefits of word-based and synset-based indexing. First, indexing is constructed for both words and synsets in the input text, and the retrieval is then performed based on either one or both of these sources of information. Our proposed method is directly inspired by these works.

The information update task shares some common properties with the IR task. In essential, the ranking models applied in our research are linear models in which features are associated with weights. These weights are learned from training data. In the learning model, some of features are similarity measures used in the IR task.

Chapter 3

Processing Methods

In this chapter, we will present the background of processing models we applied in our research. First, the information insertion task was formulated as a ranking problem. Next, we will describe two types of models: the Flat model which is trained with the standard Perceptron algorithm [13, 8], and the Hierarchical model [6] which makes use of the hierarchical decomposition of features in layers of document hierarchies. We not only describe general models for the task, but also present how these models were applied in cases of the Wikipedia insertion data and the Legal data.

3.1 The Setup: A Ranking Problem

As discussed in Chapter 2, the information insertion task can be formulated as a ranking problem. We are given an existing document and a piece of new information represented in an input sentence, we need to determine which location in the document which is most likely to be updated. In order to do that, all locations in the document at a certain level (e.g., paragraph level) are ranked by a ranking function and the location with the highest score will be chosen. The ranking function measures the relevance score between a potential location and the input sentence. There are many ways to define the ranking function. In our research we chose the supervised learning approach in which the ranking function is learned from training data.

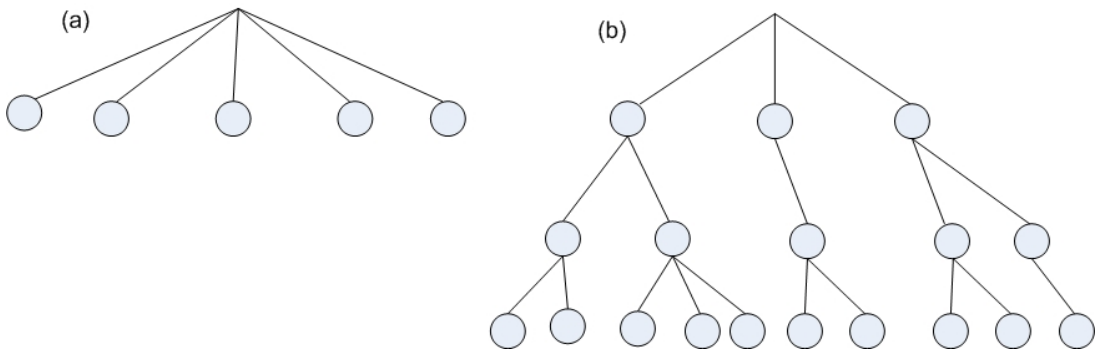


Figure 3.1: Two types of document representation. (a) A document with one layer (b) A document with multi-layers

Formally, in the information insertion task, we are given a set of training instances. Each training instance is represented by three pieces of information (s, T, ℓ) where s represents an input sentence, T is an existing document, and ℓ represents the correct insertion location of the input sentence s into the document T. Given new information, the document T can be represented as a set of potential insertion locations without considering its hierarchical structure. For example, the document contains a set of paragraphs and each paragraph is a location. The document T may be represented as a tree with multi-layers where potential insertion locations are the set of leaf nodes in the tree. We denote the set of locations of the document T in the first type and the set of leaf nodes of the document T in the second type by $L(T)$. Two types of document representations correspond to two types of models we present in the latter sections. Figure 3.1 shows two types of document representations.

In the online learning framework [6], each pair of the input sentence s and a location ℓ_i in the document is associated with a feature vector $\phi(s, \ell_i)$. As in common linear models, the ranking function $f(s, \ell)$ relies on a weight vector $\mathbf{w}$ to measure the relevance score for each given pair of the input sentence s and a potential insertion location ℓ by computing the dot product of the weight vector and the feature vector extracted from the pair. The model chooses one location among the set $L(T)$ by examining the values returned by the ranking function.

3.2 Learning methods

3.2.1 The Flat Method

Model

The model consists of a weight vector $\mathbf{w}$. For each input sentence s, the model outputs the location among the set $L(T)$ according to the following formula:

$$\hat{\ell} = \arg\max_{\ell \in L(T)} \mathbf{w}.\phi(s, \ell) \tag{3.1}$$

$L(T)$ is the set of paragraphs of the Wikipedia article T in the case of Wikipedia dataset, or the set of sections of the legal document T in the case of Legal dataset.

Training Algorithm

The task of the learning procedure in the method is reduced to learning the weight vector $\mathbf{w}$. We applied the Perceptron algorithm [13, 8] in the training phase. Algorithm 1 shows the pseudo-code for the training algorithm of the flat method.

The advantage of the online learning framework with the Perceptron algorithm is that its implementation is quite simple and it has memory-efficiency when the number of training instances is large [10].

Flat method is straightforward to understand and implement. However, it does not consider hierarchical structures of documents and relations between layers in the document tree. In a document, some paragraphs may belong to the same section, some belong to different sections. This information can be used as discriminative features. The hierarchical method was proposed to make use of hierarchical structures of documents.

Algorithm 1 Training algorithm of the Flat method

Input: Set of training instances: $(s^i, T^i, \ell^i), \ldots, (s^m, T^m, \ell^m)$

Initialize: Set $w^1 = 0$, $k = 1$

for $t = 1$ to N **do**

  for $i = 1$ to m **do**

    1. Get a new instance s^i, T^i, ℓ^i
    2. Predict $\hat{\ell}^i = \arg\max_{\ell \in L(T)} \mathbf{w}.\phi(s, \ell)$
    3. Get the correct location ℓ^i
    4. Update weight vector

    if $\hat{\ell}^i = \ell^i$ **then**

      $w^{k+1} \leftarrow w^k$

    else

      $w^{k+1} \leftarrow w^k + \phi(s, \ell^i) - \phi(s, \hat{\ell}^i)$

    end if

    $k \leftarrow k + 1$

  end for

end for

Output: w^k

3.2.2 The Hierarchical Method

Model

The main idea of the hierarchical ranking model [6] is the use of decomposition of features by defining the aggregate feature vector of each leaf node in the document tree. The aggregate feature vector of a leaf node was defined to be the sum of all features at the upper layers.

Formally, for each sentence and a document T, the feature vectors set of all nodes in the document tree is denoted by $\phi(s, n) : n \in T$. Denote the set of leaf nodes by $L(T)$ and the path from the root of the tree to a node n by $P(n)$. The aggregate feature vector

associated with a leaf node ℓ is computed by the following formula.

$$\Phi(s, \ell) = \sum_{n \in P(\ell)} \phi(s, n) \tag{3.2}$$

The decoding method of the hierarchical is similar to the decoding method of the flat method except the use of aggregate vectors for leaf nodes in the tree.

$$\hat{\ell} = \arg\max_{\ell \in L(T)} \mathbf{w}.\phi(s, \ell) \tag{3.3}$$

The hierarchical method can be applied even for dataset in which number of layers in document hierarchies varies from document to document. However, to be simple for implementation and evaluation, in [6], a Wikipedia article was assumed to be divided into sections and then paragraphs. For legal documents, we assumed that there are three layers for every document hierarchy: sections as leaf nodes, the intermediate upper layer of sections, and the document as the root. Hereafter, we use the term *chapter-layer* for the intermediate upper layer of sections.

Training Algorithm

The training algorithm was based on the heuristic observation: if the model incorrectly predicts at a certain layer, then its children layers will not be considered. This heuristic was incorporated in the training algorithm that only weights of features at the split point between predicted path and the true path are updated. The update rule for each round is defined as below.

$$\mathbf{w} \leftarrow \mathbf{w} + \phi\big(s, P(\ell)^{i^*+1}\big) - \phi\big(s, P(\hat{\ell})^{i^*+1}\big) \tag{3.4}$$

where $\hat{\ell}$ denotes the predicted leaf node, and ℓ is the correct leaf node; $P(\ell)^i$ denotes the i^{th} node on the path from the root to ℓ, and i^* is defined as the depth of the lowest common ancestor of ℓ and $\hat{\ell}$.

Figure 3.2 illustrates the idea of the heuristic update rule, and Algorithm 2 showes the pseudo-code of the training algorithm of the hierarchical method.

It was assumed that each Wikipedia article is composed of sections and each section is divided into paragraphs. If the model fails to predict correct section, only weights of

Algorithm 2 Training algorithm of the Hierarchical method

Input: Set of training instances: $(s^i, T^i, \ell^i), \ldots, (s^m, T^m, \ell^m)$

Initialize: Set $w^1 = 0$, $k = 1$

for $t = 1$ to N **do**

  for $i = 1$ to m **do**

    1. Get a new instance s^i, T^i, ℓ^i
    2. Predict $\hat{\ell}^i = \arg\max_{\ell \in L(T)} \mathbf{w}.\Phi(s, \ell)$
    3. Get the correct location ℓ^i
    4. Update weight vector

    if $\hat{\ell}^i = \ell^i$ **then**

      $w^{k+1} \leftarrow w^k$

    else

      $j^* \leftarrow \max\{j : P(\ell)^j = P(\hat{\ell})^j\}$

      $w^{k+1} \leftarrow w^k + \phi\big(s, P(\ell)^{i^*+1}\big) - \phi\big(s, P(\hat{\ell})^{i^*+1}\big)$

    end if

    $k \leftarrow k + 1$

  end for

end for

Output: w^k

section features are updated. This heuristic rule is the same to legal documents with two layers: chapter-layers and section-layer.

In the training algorithm of the hierarchical method, if we do not use the heuristic update rule in the training algorithm, the model will become the Flat method with additional features from upper layers of leaf nodes in the document hierarchy. In experiments, we compared the method of using the heuristic update rule with the setting without using the rule.

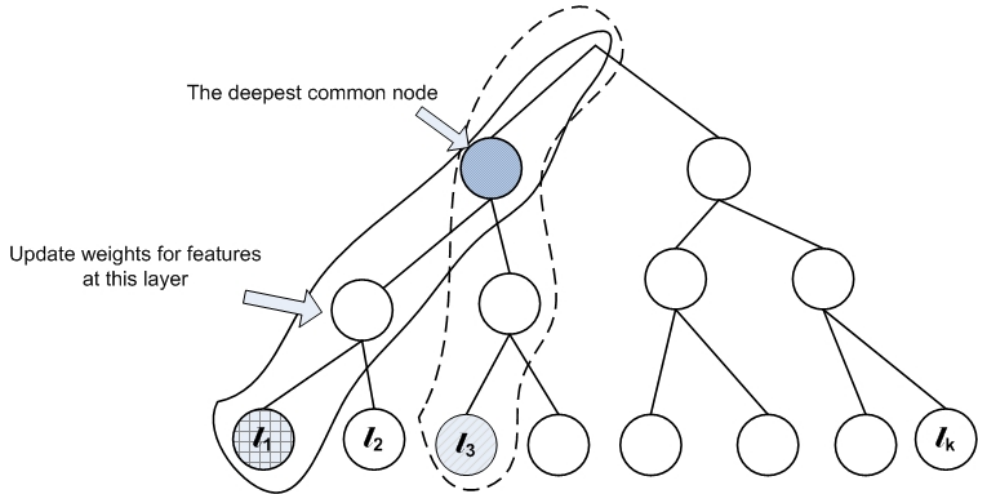


Figure 3.2: Heuristic update rule in the training algorithm of the Hierarchical method. l_1 is the correct location, l_3 is the predicted location.

3.3 Feature Extraction in Two Methods

There are some important points in the hierarchical method. First, the hierarchical method makes use of features at multi layers in a document hierarchy while features in the basic Flat method are extracted at only one level. The second point is the heuristic update rule in the training algorithm.

In essential, the key point that makes the hierarchical method different from the Flat method is the way features are extracted at different levels. Generally, features are extracted hierarchically. It means that features of a node at a certain layer are extracted within its parent node to distinguish this node from its sibling nodes. To be clearer, we give an example of a document with three layers: document, section level and paragraph level (the document is the root of the tree). At the section level, we compute the TF-IDF weighted cosine similarity between an inserted sentence with a section within the whole document. At the paragraph level, the TF-IDF weighted cosine similarity between an inserted sentence with a paragraph is computed within the section to which the paragraph belongs instead of the document as a whole. The TF-IDF weighted cosine similarity at paragraph level is actually for the local ranking among paragraphs within each section.

Chapter 4

Semantic Features based on Word Clustering

In this chapter, we will present our method of incorporating semantic features derived from Word Clusters into the learning model for the information update task. First, we will give a brief background of Word Clustering and its utility in natural language processing tasks. Then, we will describe how we computed and used cluster-based features in details. Some issues that we must consider in our method will be discussed below.

4.1 Background in Word Clustering

4.1.1 Benefits of Word Clustering in NLP

Word clustering is a process of assigning words to classes [5]. Each class contains words which are semantically or syntactically similar. For example, the word *Thursday* is very much like the word *Friday* due to their function in expressing a day in a week, so they should be assigned to the same word class.

In lexical semantics, there are two types of the semantic relation, the semantic similarity and the semantic relatedness [39, 16]. Two words are semantically similar if they appear in similar contexts and they may be substituted for another. For example, in the context "I

met the chairman", the word *chairman* can be replaced by the word *president*, and these two words can be considered to be semantically similar. In the latter type of semantic relation, two words are considered to be semantically related if they significantly co-occur within the same context. For instance, two words *cut* and *knife* are semantically related. Two kinds of semantic relations have been seen as important concepts in natural language processing. Word clustering is a technique for assigning sets of words into classes of semantically similar words and it can capture both types of word relations. Thus, it is becoming a major technique used in many natural language processing tasks.

In some Natural Language Processing tasks, word clustering can be used to tackle the problem of data sparseness by providing a lower-dimensional representation of words. For instance, in the well-studied text classification task which is to assign a document to one or more categories, words are typically used as features in discriminative learning algorithms [2, 4]. For the case of text collections with a very large size vocabulary, feature vectors have very high dimension, and they are usually spare. With a good word clustering, some words which are semantically related can be merged without hurting the classification performance, and then the number of features needed for text classification can be reduced.

The method of incorporating features based on word clustering into a discriminative learning framework has been previously explored by Miller et al. [27] in the Named-Entity Recognition task. The success obtained in the work of Miller has inspired many other researches. Liang [24] also used word cluster-based features in Named-Entity Recognition and Chinese Word Segmentation tasks. In the Dependency Parsing task, an important topic in natural language processing, Koo et al. [21] demonstrated the effectiveness of additional features that incorporate word clusters for parsing syntactic structure. The accuracy of dependency parsing with cluster-based features in the cases of English and Czech improved over the baseline accuracy.

In the Information Retrieval task, word clustering can be used as an automatically generated similarity thesaurus for query expansion. A query can be expanded by adding all terms in the word classes that contain the query terms [34].

In our research, we used word clusters as an intermediate word representation when computing semantic text similarity. We expect that combining word clusters with surface forms of words can exploit semantic relatedness of words better than using the words themselves.

4.1.2 Word Clustering Algorithms

Word clustering algorithms are typically parted according to the kind of semantic similarity they take into account, the semantic similarity and semantic relatedness. For the first kind, the semantic similarity of words is computed either based on taxonomical relationship of words in a hierarchically structured lexical resource such as WordNet or based on their distributions in contexts which they appear [39]. The latter approach takes into account the co-occurrence of words with others in a large text corpus.

In our research, we used the English word clusters computed by the Brown Word Clustering algorithm [5], a statistical algorithm for assigning words to classes based on the frequency of their co-occurrence with other words in a large text data. Following is a brief description of the algorithm.

4.1.3 Brown Word Clustering Algorithm

Brown word clustering algorithm received a vocabulary V of words to assign to classes and a text corpus as input. In the initial step, each word in the vocabulary V is assigned to a distinct class, and the average mutual information between adjacent classes is computed. The algorithm then repeatedly merges the pairs of classes for which the loss in average mutual information is the least. If C classes are required, $V - C$ merges need to be performed. The output of the algorithm is a hierarchical clustering of words represented in a binary tree, where each leaf node has a word and, each word occupies in only one leaf node. Each internal node at a certain layer is a word cluster which contains all words in the sub-tree derived from that node. A word in the vocabulary V can be assigned to a binary string by traversing the path from the root to its leaf node, assigning a bit 0

1001111011011	economic-consulting
1001111011011	investment-advisory
1001111011011	management-consulting
1001111011011	financial-planning
1001111011011	asset-management
...	
100111101011	electronics-parts
100111101011	vending-machine
100111101011	engine-overheating
100111101011	computer-peripherals
100111101011	industrial-electronics
100111101011	sewing-machine
...	
10100100010	legislator
10100100010	policeman
10100100010	caller
10100100010	soldier
10100100010	detective
10100100010	composer
10100100010	poet

Table 4.1: Examples of word clusters. Words having the same binary string representation belong to the same cluster

for branches in the left and a bit 1 for branches in the right. The Brown word clustering generates a hard clustering in which a word belongs to only one word class. Figure 4.1 illustrates a binary tree which represents a Brown word clustering hierarchy.

The advantage of the Brown word clustering is that it only requires a raw text data corpus which is available in various sources such as the internet. However, the computational complexity of the algorithm is $O(k^3)$ in [5] and $O(k^2)$ in the implementation of Liang [24],

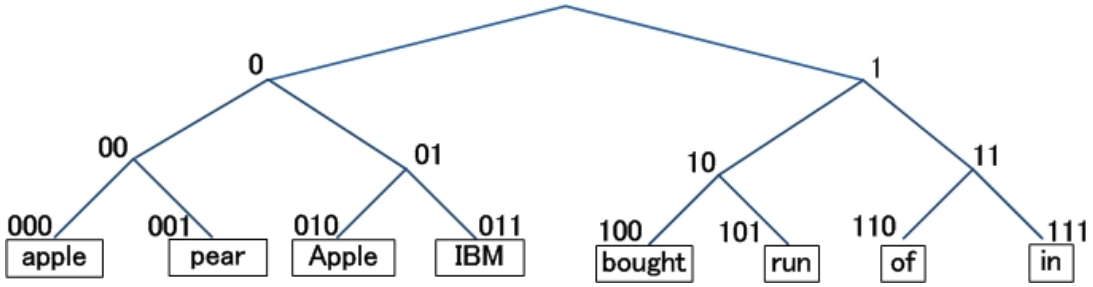


Figure 4.1: An example of a Brown word-cluster hierarchy (Koo et al., 2008)

here k is the number of clusters. With large input text corpus, the computational time for computing word clusters is quite long.

To conduct experiments in this paper, we used the English word clusters corpus from [21] including 1000 word clusters. The Liang implementation [24] of the Brown algorithm was used to obtain those word clusters. Table 4.1 provides some example binary strings.

4.2 Extracting Semantic Features based on Word Clustering

4.2.1 Baseline Topical Overlap Features

In the task of adding new information into an existing document, the topical overlap between an input sentence and sentences of each candidate insertion location is one of the important features to preserve the text coherence of the original text. The topical overlap features become even more important in the case of legal dataset because most features extracted for legal data are based on topical overlap.

In [6], the topical overlap features were computed using the TF-IDF weighted cosine similarity and word overlap between the input sentence and the set of sentences in each potential location. Table 4.2 shows some baseline overlap features.

Paragraph level baseline topical overlap features
The number of sentences in p which shared non-stop-words/nouns/proper nouns/verbs with *sen*
TF score between p and *sen* based on non-stop-words/nouns/proper nouns/verbs
TF-IDF score between p and *sen* based on non-stop-words/nouns/proper nouns/verbs
Section level baseline topical overlap features
The number of sentences in *sec* which shared non-stop-words/nouns/proper nouns/verbs with *sen*
TF score between *sec* and *sen* based on non-stop-words/nouns/proper nouns/verbs
TF-IDF score between *sec* and *sen* based on non-stop-words/nouns/proper nouns/verbs

Table 4.2: List of some baseline topical overlap features at paragraph level (given an insertion sentence *sen*, and a paragraph p)

4.2.2 Cluster-based Topical Overlap Features

The simple lexical matching approach to compute baseline topical overlap features was based on the "bag of words" assumption, using the surface forms of words in the input sentence and each candidate location. However, words in the input sentence may not appear in the locations, so the semantic relations between semantically related words are not exploited. In order to exploit these relations, we propose a method of using word clusters as intermediate word representations to obtain semantic features as follow.

Assume that we are given word clusters obtained by running the Brown word clustering algorithm on a text corpus. Since the Brown algorithm produces hierarchical clusters, the question is which layer in the hierarchy we should/will choose to generate word classes. We will choose the layer in the hierarchy so that the number of word clusters is large enough. After picking word clusters from the hierarchy, we obtain a set W which contains words in the vocabulary V along with their classes. Word classes are represented as binary strings as discussed above. Note that generated word clustering is a hard clustering, so each word has only one binary string representation. For concreteness, we use the notation $\pi(w)$ to

denote the mapping from each word w to its binary string representation. In experiments, if w is not in lowercase form, we heuristically add the binary string representation of w in lowercase to $\pi(w)$. It means that the mapping $\pi(w)$ of a word w which is not in lowercase contains two binary strings, the first is the binary string of w, another one is the binary string of w in lowercase.

For each pair of an input insertion sentence s and a node n which is a set of sentences (n can be a paragraph, or section in the Wikipedia insertion data or a section, or a chapter in the legal data), we performed the following three steps:

Step 1: Obtain the binary string representation for each word in s and n by the mapping π. Words which are not included in the vocabulary V will be assigned to a special value *null*.

Step 2: Compute the text similarity of two text segments s and n based on their binary string representations.

Step 3: Incorporate text similarity scores obtained in above steps into the learning model as additional features.

In the step 2, we need to determine the text similarity function $f(s, n)$ to measure similarity of two text segments s and n. In our research, we used some kinds of text similarity functions as follows.

TF and TF-IDF Weighted Cosine Similarity

The first text similarity function used in the step 2 is the TF and TF-IDF weighted cosine similarity function. TF (term frequency) and TF-IDF (term frequency-inverse document frequency) are weighting schemes often used in information retrieval and text mining. While the term frequency measures the importance of a term in a particular document, the inverse document frequency measures the general importance of a term in the whole text collection. The TF and TF-IDF weighted cosine similarity are generally used to compute similarity of two text segments after performing term weighting on both of them.

After having binary string representations of words in s and n, the system will perform

weighting on these binary strings with TF and TF-IDF weighting scheme. Like in [6, 7], we in turn computed TF and TF-IDF weighted cosine similarity based on binary string forms of all words excluding stop words, binary string forms of nouns, proper nouns, verbs. We did not take into account adjectives which do not contribute to topics of a text. In effect, our method is somewhat similar to the method of semantic indexing using WordNet synsets [15, 25].

The Lexical Matching Function

The lexical matching function measures the lexical-based semantic overlap of two text segments. In our task, two text segments are the sentence s and a certain node n in the document tree. It is a score based on matching each word in s with words in n. The lexical matching score is the percentage of words in s appearing in n.

The lexical matching function of two text segments s and n is computed by the following equation.

$$LexMatch(s, n) = \frac{\sum_{w \in s} \mu(w, n)}{|s|} \tag{4.1}$$

where $|s|$ denotes the number of words in the sentence s. To incorporate word clusters, we define the $\mu(w, d)$ function of a word w and a "set of words" d as follows:

$$\mu(w, d) = \begin{cases} 1 & \text{if} \quad \exists w^* \in d \quad \text{so that} \quad w^* = w \quad \text{or} \quad \pi(w) \cap \pi(w^*) \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

The equation 4.2 can be interpreted that a word w is said to be matched with a "set of words" d if w appears in d or there is a word in the same cluster with w in the set d.

The Average Jaccard Similarity Function

The average Jaccard similarity function between a sentence s and a node n is computed by averaging out the Jaccard similarity scores [19] of all sentences in n with s.

$$AvgJacSim(s, n) = \frac{\sum_{v \in n} JacSim(s, v)}{|n|} \tag{4.3}$$

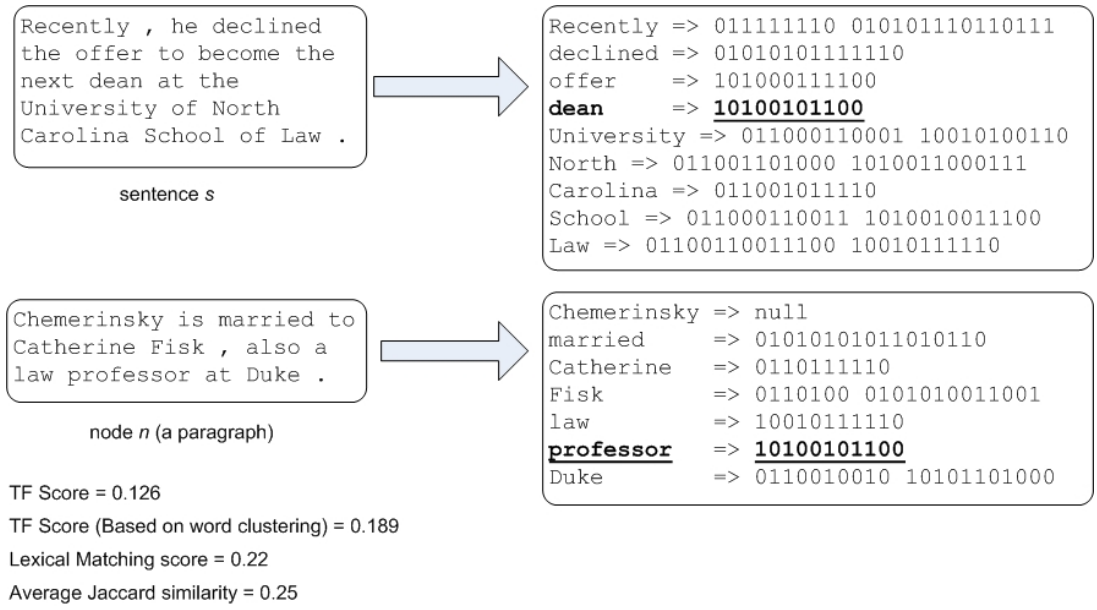


Figure 4.2: How text similarity functions are computed

Here, v is any sentences in the node n, $|n|$ represents the number of sentences in n. The Jaccard similarity function of a sentence pair is computed by the following equation.

$$JacSim(s_1, s_2) = \frac{\sum_{i \in s_1} \mu(i, s_2) + \sum_{j \in s_2} \mu(j, s_1)}{|s_1| + |s_2|} \tag{4.4}$$

In the Equation 4.4, the function μ is the same as in the lexical match function 4.1.

The Figure 4.2 illustrates how text similarity functions are computed.

In the example in the Figure 4.2, our proposed method can capture the semantic relations between two words *dean* and *professor*.

Semantic features based on Word Clustering are computed for all layers of each document tree. Table 4.3 gives the list of topical overlap features based on Word Clustering at the paragraph level in the case of Wikipedia insertion dataset.

Features at the upper levels of the document hierarchy are computed in the similar way.

Paragraph level cluster-based features
TF score between p and *sen* based on binary string representations of non-stop-words
TF score between p and *sen* based on binary string representations of nouns
TF score between p and *sen* based on binary string representations of proper nouns
TF score between p and *sen* based on binary string representations of verbs
TF-IDF score between p and *sen* based on binary string representations of non-stop-words
TF-IDF score between p and *sen* based on binary string representations of nouns
TF-IDF score between p and *sen* based on binary string representations of proper nouns
TF-IDF score between p and *sen* based on binary string representations of verbs
Lexical matching score of *sen* with p based on word clusters
Average Jaccard similarity score of *sen* and p based on word clusters

Table 4.3: List of some sample features based on word clusters (given an insertion sentence *sen*, and a paragraph p)

4.3 Some Discussions

In our method of extracting semantic features that incorporate word clusters, there are some issues we must consider.

1. The raw text corpus from which we extract word clusters should be large enough to cover the documents in our corpus.

2. The domains of the raw text corpus should be close to the domains of documents in our corpus. The co-occurrence frequency of words with others may vary in different domains; therefore, it affects the extracted word clusters.

3. We must consider the level in hierarchical word clustering, from which we pick word clusters. In other words, that is the number of word clusters we use in our method. Experiments in our research were conducted using English word clusters data from [21] with 1000 word classes.

Word clustering can capture semantic relations of words which are semantically similar, but it provides no word-to-word similarity measure. In fact, word-to-word similarity metrics are important to know if a word is more semantically similar with a certain word than with others in a particular context, and measuring text-to-text similarity can make use of word-to-word similarity.

Chapter 5

Experiments and Results

In this chapter, we will present our experiments and results on two datasets: Wikipedia insertion data and Legal data. The effect of using the hierarchical ranking models for the information insertion task was reported in [6] on the Wikipedia insertion data. However, what will happen if we apply these model for the Legal data? In our experiments, first, we investigate processing methods on both of datasets. Second, we conduct experiments with cluster-based semantic features to demonstrate the effect of using word cluster-based features for the information insertion task.

5.1 Data Preparation

5.1.1 Wikipedia Insertion Dataset

As introduced in Chapter 2, in experiments, we obtained Wikipedia insertion data from [6]. The Wikipedia insertion data consists of 4051 insertion/article pairs derived from 1503 Wikipedia articles in the category "Living people". On average, an article in the dataset has 32.9 sentences, organized in 3.61 sections and 10.9 paragraphs [7].

Each document in the Wikipedia insertion data is composed of sections, and each section is composed of paragraphs. Thus, the document hierarchy has three layers: document-section-paragraph. Paragraphs in a document are leaf nodes in the document tree.

Title 1	General Provisions
Title 2	The Congress
Title 3	The President
Title 4	Flag and Seal, Seat of Government, and the States
Title 5	Government Organization and Employees
...	
Title 47	Telegraphs, Telephones, and Radiotelegraphs
Title 48	Territories and Insular Possessions
Title 49	Transportation
Title 50	War and National Defense

Table 5.1: List of some Titles in the United States Code

5.1.2 Legal Dataset

We have built the Legal dataset from legal documents of the United States Code data. The United States Code (USC) is a compilation and codification of general and permanent federal law of the United States [41]. It is divided by broad subjects into 50 titles and published by the Office of the Law Revision Counsel of the U.S. House of Representatives.

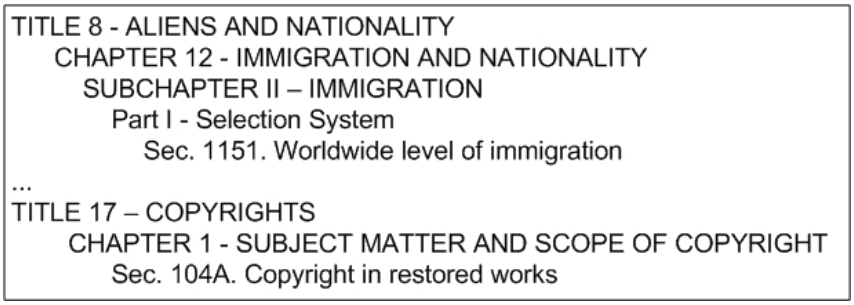


Figure 5.1: Example of document structures of the Title 8 and Title 17.

AMENDMENTS
2005 - Pub. L. 109-9 inserted definition of "motion picture Exhibition facility" after definition of "Motion pictures".
2004 - Pub. L. 108-419 inserted definition of "Copyright Royalty Judge" after definition of "Copies".
2002 - Pub. L. 107-273, Sec. 13210(5)(B), transferred definition Of "Registration" to appear after definition of "publicly".
Pub. L. 107-273, Sec. 13210(5)(A), transferred definition of "computer program" to appear after definition of "compilation".
2000 - Pub. L. 106-379, Sec. 2(a)(2), in definition of "work made for hire", inserted after par. (2) provisions relating to considerations and interpretations to be used in determining whether any work is eligible to be considered a work made for hire under par. (2).
Pub. L. 106-379, Sec. 2(a)(1), in definition of "work made for Hire", struck out "as a sound recording," after "motion picture or other audiovisual work," in par. (2).
1999 - Pub. L. 106-113, which directed the insertion of "as a Sound recording," after "audiovisual work" in par. (2) of definition relating to work made for hire, was executed by making The insertion after "audiovisual work," to reflect the probable intent of Congress.
Pub. L. 106-44, Sec. 1(g)(1)(B), in definition of "proprietor", Substituted "For purposes of section 513, a 'proprietor' " for "A 'proprietor' ".
...

Figure 5.2: Amendment part of the section Sec. 101. Definitions; Chapter 1 - Subject matter and scope of copyright; Title 17 - Copyrights

Legal documents are different from general texts in terms of structures. We have analyzed the legal documents in the United States Code. In general, legal documents of the US Code have characteristics as follows:

- Legal documents have highly hierarchical structures in which a document is divided into subdivisions with various layers. For example, a document may optionally be divided into subtitles, parts, subparts, chapters, and subchapters, etc. An example of structures of legal documents is showed in figure 5.1.
- Not all documents use the same series of subdivisions, and they may arrange them in different orders.
- All documents have sections as their basic coherent units, though sections are often divided into (from the largest to the smallest) subsections, paragraphs, subparagraphs, clauses, sub-clauses, items, and sub-items.

In each document of the U.S Code, there are amendment parts after every section. Amendment parts record revisions on the corresponding section. Therefore, the data for

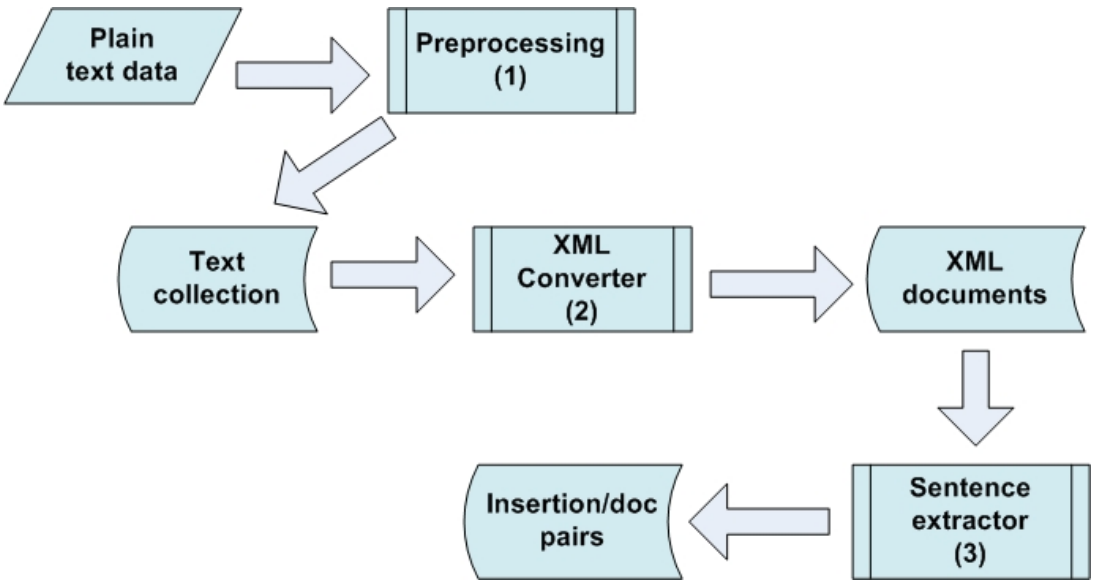


Figure 5.3: Phases of constructing Legal dataset

the insertion task can be built from these parts. Figure 5.2 provides the amendment part after the Sec. 101., Chapter 1 of the Title 17.

The amendment parts do not record the statutes before and after changes. Thus, it is difficult to automatically obtain data for the insertion task from amendment parts. For the purpose of evaluating methods proposed in our research, we decided to build data in an automatic way.

Figure 5.3 illustrates phases of constructing the Legal dataset. There are three phases in constructing data: preprocessing phase, XML converting, and sentences extracting.

Preprocessing Phase

In the preprocessing phase, in each document we remove redundant information other than contents of statutes, such as tables and contents, information about amendments, etc. Next we perform POS tagging for all documents using Stanford POS tagger tool [37, 38].

```
<title id='17'>
  <name>  TITLE 17 – COPYRIGHTS </name>
  <chapter id='1'>
     <name> CHAPTER 1 - SUBJECT MATTER AND SCOPE OF COPYRIGHT </name>
     <sec id='101'>
        <name> SEC. 101. DEFINITIONS </name>
        <statute>
           <sen id='0'>Except_IN as_IN otherwise_RB provided_VBN in_IN this_DT Title_NN ,_, as_IN
           used_VBN in_IN this_DT title_NN ,_, the_DT following_JJ terms_NNS aNd_CC their_PRP$
           variant_JJ forms_NNS mean_VBP the_DT following_NN :_:</sen>
           <sen id='1'>An_DT ``_`` anonymous_JJ work_NN "_" is_VBZ a_DT Work_NN on_IN
           the_DT copies_NNS or_CC phonorecords_NNP of_IN which_WDT no_DT natural_JJ
           person_NN is_VBZ identified_VBN as_IN author_NN ._.</sen>
           ...
        </statute>
     </sec>
     <sec id=' 104A' >
      ...
     </sec>
     ...
  </chapter>
  <chapter id='2'>
     <name>  CHAPTER 2 - COPYRIGHT OWNERSHIP AND TRANSFER </name>
     <sec id='201'>
        <name> SEC. 201. OWNERSHIP OF COPYRIGHT </name>
        <statute>
           <sen id='1288'>-_: Copyright_NN in_IN a_DT work_NN protected_VBN under_IN this_DT
           title_NN vests_VBD initially_RB in_IN the_DT author_NN or_CC authors_NNS of_IN the_DT
           work_NN ._.</sen>
           <sen id='1289'>The_DT authors_NNS of_IN a_DT joint_JJ work_NN are_VBP coowners_NN
           of_IN copyright_NN in_IN the_DT work_NN ._.</sen>
           <sen id='1290'>(_( b_NN )_) Works_NNP Made_NNP for_IN Hire_NNP ._.</sen>
           <sen id='1291'>-_: In_IN the_DT case_NN of_IN a_DT work_NN made_VBN for_IN hire_NN
           ,_, the_DT employer_NN or_CC other_JJ person_NN for_IN whom_WP the_DT work_NN
           was_VBD prepared_VBN is_VBZ considered_VBN the_DT author_NN for_IN purposes_NNS
           of_IN this_DT title_NN ,_, and_CC ,_, unless_IN the_DT parties_NNS have_VBP
           expressly_RB agreed_VBN otherwise_RB in_IN a_DT written_VBN instrument_NN
           signed_VBN by_IN them_PRP ,_, owns_VBZ all_DT of_IN the_DT rights_NNS
           comprised_VBN in_IN the_DT copyright_NN ._.</sen>
           ...
        </statute>
     </sec>
     ...
  </chapter>
  ...
</title>
```

Figure 5.4: A part of Title 17 in XML format

XML Converter

Documents in the U.S Code have hierarchical structures with multi-levels; and intermediate levels vary in both number and sequence across documents. Therefore, we need to convert legal documents into a text format which can represent hierarchical structures, and allow us to easily access contents in the hierarchy.

We choosed the XML format (Extensible Markup Language) to represent hierarchical structures of texts. Since XML allows users to define the markup-elements, hierarchical structures are clearly represented. By making use of symbolic characteristics of plain text version of the U.S Code, we successfully constructed 49 XML files from 49 text files corresponding to 49 documents in the U.S Code (title 34 was repealed, so it was not in our collection). The Figure 5.4 shows a part of an example XML document.

Sentence Extractor

In this phase, we have built the dataset automatically by randomly extracting one sentence from each legal document and recording the document before and after removing the sentence from it. We repeatedly performed this process on documents until we obtained reasonable amount of data. Our dataset consists of extracted sentence/document pairs built by this method. We eliminated pairs whose insertion sentences are too short or too long, and pairs whose insertion sentences belonging to sections of only one sentence. In cases where multiple sentences were extracted from the same document, they are treated independently to each other. Totally, we obtained 1812 insertion sentence/document pairs from 18 legal documents. Legal documents in our dataset are very long documents. Average document has 1472.4 sentences, organized in 141.9 sections.

The main advantage of our method in building dataset is that it saves time and human efforts. The disadvantage is that the dataset seems not to be very natural, and it lacks of information about user habits in editing documents. Since in the insertion task for Legal domain, we only take into account the topical overlap between new information and potential sections in a document to place new information, so the dataset built by this way is sufficient for evaluating our proposed method.

	docs	insertion/doc pairs	instances (train data)	instances (test data)	average places per doc	average sentences per doc
Wikipedia	1503	4051	3240	811	10.9	32.9
Legal	18	1812	1450	362	141.9	1472.4

Table 5.2: Some statistic information of datasets. In Wikipedia data, the smallest coherent unit is a paragraph. In Legal data, the smallest coherent unit is a section.

5.2 Experimental Setting

5.2.1 Dataset

For each insertion dataset, we use 80% data for training, and 20% data for testing. Thus, for Wikipedia data, we used 3240 insertion/article pairs for training, and 811 pairs for testing. For the Legal data, training data consists of 1450 insertion/document pairs, and testing data consists of 362 insertion/document pairs.

5.2.2 English Word Clusters

We used English Word cluster set from [21], which was exploited successfully for the Dependency Parsing task. The word clusters were derived from BLLIP corpus including 43 million words of Wall Street Journal text. There are 1000 clusters and 316710 word types in total.

5.2.3 Features

Wikipedia Data

On Wikipedia data, we used the feature set as reported in [7] and additional word cluster-based features described in the Chapter 4 of this thesis.

The baseline features set for Wikipedia contain three types of features: lexical, positional, and temporal features. The lexical features capture the topical overlap of an inserted sentence and a paragraph in a document while positional features aim to capture

Section level features
The number of sentences in *sec* which shared non-stop-words/nouns/proper nouns/verbs with *sen*
Whether *sen* and the title of *sec* shared noun words
TF score between *sec* and *sen* based on non-stop-words/nouns/proper nouns/verbs
TF-IDF score between *sec* and *sen* based on non-stop-words/nouns/proper nouns/verbs
TF score between *sec* and *sen* based on binary string representation of non-stop-words/nouns/pronouns/verbs
TF-IDF score between *sec* and *sen* based on binary string representation of non-stop-words/nouns/pronouns/verbs
Lexical matching score of *sen* with *sec* based on word clusters
Average Jaccard similarity score of *sen* and *sec* based on word clusters

Table 5.3: Some section level features for Legal dataset (*sen* is an input sentence, and *sec* is a section)

user preferences when adding new information into the body of a document. For instance, users of Wikipedia tend to add new information at the end of a section than its beginning. The third type of features is the temporal feature obtained from the observation that in articles of the category "Living people", events about an individual are often organized chronologically.

On Wikipedia data, the features are computed in two levels: the section level and the paragraph level. The topical overlap features in section level are computed in a similar way with features in the paragraph level.

Legal Data

Since Legal dataset have been built in a synthetic way, the positional and temporal features were not used. Therefore, on the Legal data, only lexical features and word cluster-based features were extracted. Table 5.3 lists some features at section level in experiments on the Legal data.

Generally, structures of legal documents are more complicated than structures of Wikipedia articles. There are more layers in a document hierarchy, and intermediate layers above sections vary from document to document. In principle, features at all level can be computed, but for the evaluation purpose, we only consider features at two levels, the section level and the intermediate upper level of the section level. We call the intermediate upper level of the section level by the chapter level. The features at the chapter level are computed in a similar way.

5.2.4 Evaluation Measures

Wikipedia Data

We used the same evaluation measures as in [6] : a) insertion accuracy and b) the tree distance between the predicted and the true location of the inserted sentence. Insertion accuracy is the percentage of matches between predicted location of insertion and the true placement; and tree distance is defined as the length of the path through the tree which connects the predicted and the true paragraph positions. Shorter tree distance corresponds to the better performance.

Legal Data

We used two evaluation measures in experiments on Legal data as follows.
a) *Accuracy of choosing sections* is the percentage of correct predictions, and computed by the following equation.

$$\text{Accuracy of choosing sections} = \frac{\text{Number of correct predictions}}{\text{Total number of insertion/doc pairs in data}} \tag{5.1}$$

b) *N-best accuracy*: A prediction will be judged correct if the correct section is in the top N sections returned by the ranker. In experiments, we choose $N = 5$ and $N = 10$.

5.2.5 Processing Methods to Evaluate

The main part of our thesis is the proposed method of incorporating cluster-based features to the information update task, so baselines in our experiments are processing methods without using cluster-based features. We investigated methods as follow.

Flat method is the method in which the model is trained by standard Perceptron algorithm, using only features in the leaf node level of a document hierarchy.

Hier-1 method is the hierarchical method without using the heuristic update rule in the training algorithm.

Hier-2 method is the hierarchical method using the heuristic update rule in the training algorithm.

In [6], supervised learning methods were proposed, but there is no comparison between the supervised approach and the unsupervised approach. The comparison between two approaches is interesting; especially in the case of Legal data in which the features set is not very abundant. Thus, we conducted experiments with the unsupervised method using the TF-IDF weighted cosine similarity between an inserted sentence and a location as the ranking function.

5.3 Results

5.3.1 Effect of Using the Hierarchical Ranking Models

The effect of using the Hierarchical ranking model was reported in [6]. Experiments on the Wikipedia insertion data showed that the Hierarchical ranking model with the heuristics update rule significantly outperformed other methods.

	Section (%)	**Paragraph (%)**	**Tree distance**
Unsupervised (TF-IDF)	53.5	27.3	2.38
Flat	57.9	31.4	2.21
Hier-1	58.9	34.2	2.13
Hier-2	59.8	38.3	2.04

Table 5.4: Results on Wikipedia dataset with baseline features

	Section (%)	**5-best (%)**	**10-best (%)**
Unsupervised (TF-IDF)	41.4	75.4	85.0
Flat	47.8	76.2	85.3
Hier-1	50.9	81.6	89.5
Hier-2	50.9	81.8	89.1

Table 5.5: Results on Legal dataset with baseline features (5-fold cross validation)

We investigated the processing methods on the new legal dataset, comparing the performance of different methods to confirm the effect of using the Hierarchical ranking model for the information update task. In experiments of this section, we only used baseline features. For legal dataset, we perform 5-cross validation. First, we divided data into 5 subsets. We in turn used each subset as validation set, and used four remaining subsets as training set. Finally, we computed average accuracy of five iterations.

Table 5.4 and 5.5 shows the performance of methods on the Wikipedia insertion dataset and the Legal dataset respectively.

The results indicate that, on the Legal dataset, since the extracted features mainly aim to capture text similarity the unsupervised method obtained quite good performance relative to supervised methods, 41.4% accuracy of choosing sections. It implies that for some kinds of data, the unsupervised method is a promising method if a good ranking function can be designed.

On two datasets, Hierarchical methods outperform the unsupervised method and the Flat method. However, while the Hier-2 method significantly improved over other methods

		Section (%)	Paragraph (%)	Tree distance
Baseline	Flat	57.9	31.4	2.21
	Hier-1	58.9	34.3	2.13
	Hier-2	59.8	38.3	2.04
Setting-1	Flat	59.5	36.6	2.08
	Hier-1	59.3	35.1	2.11
	Hier-2	59.1	38.5	2.05
Setting-2	Flat	59.5	36.2	2.08
	Hier-1	59.8	36.4	2.07
	Hier-2	60.2	40.4 (+2.1)	1.99

Table 5.6: Results on Wikipedia dataset of three settings

on Wikipedia data, the performance of the Hier-1 method and the Hier-2 method were almost the same on the new Legal data. It can be explained that on the Legal data, there are not many features at section level to distinguish sections within a chapter (only TF-IDF weighted cosine similarity scores were used). Another reason is that in Hierarchical methods, chapters in the chapter level are not real chapters in documents. They are very different in length, number of sections, etc.

5.3.2 Effect of Using Semantic Features based on Word Clustering

To demonstrate the effect of using semantic features based on Word clustering, we conducted experiments on two datasets with three following settings.

In the **Baseline** setting, we only used baseline features set without cluster-based features.

In the **Setting-1**, we use lexical cluster-based features instead of corresponding baseline lexical features. One cluster-based feature is said to be corresponding with a certain baseline feature if they are computed in the similar way except that we combine the binary string forms with surface forms of words instead of using the words themselves.

		Section (%)	5-best (%)	10-best (%)
Baseline	Flat	47.8	76.2	85.3
	Hier-1	50.9	81.6	89.5
	Hier-2	50.9	81.8	89.1
Setting-1	Flat	46.0	73.8	83.5
	Hier-1	49.3	80.5	88.1
	Hier-2	49.6	79.8	88.0
Setting-2	Flat	49.5	80.0	87.0
	Hier-1	52.0	83.4	89.9
	Hier-2	52.3 (+1.4)	83.0	90.1

Table 5.7: Results on Legal dataset of three settings (5-fold cross validation)

In the **Setting-2**, we combine word cluster-based features with all baseline features. Table 5.6 and 5.7 shows the performance of processing methods in three settings.

The results in the **Setting-1** were not stable on two datasets. On the Wikipedia dataset, the performance of the Flat method and Hier-1 in **Setting-1** improved comparing to the **Baseline** setting, but the performance of the Hier-2 method slightly decreased. On Legal data, the performance of methods in the **Setting-1** did not outperform the **Baseline** setting. It may be caused by the out of domain problem. In our method, a word is substituted by its binary string if the word appears in the vocabulary of Word clusters corpus. If the vocabulary is not large enough to cover all words in the dataset, many words will be omitted, and it leads to the information loss. Especially, in the case of Legal dataset, there are many legal terminologies which are not included in the vocabulary of the chosen English word clusters.

The **Setting-2** which combines word cluster-based features with baseline features overcome problems in the **Setting-1**. All methods in the **Setting-2** got better performance than other settings, and the Hier-2 method which used the heuristic update rule in the training algorithm obtained the best performance, 40.4% accuracy of choosing paragraphs on Wikipedia dataset and 52.3% accuracy of choosing section on the Legal dataset.

Chapter 6

Conclusion

The task of updating information is a new and challenging task in Natural Language Processing, especially in Legal domain where legal documents must be updated consistently. Inspired by the work in [6], our research addressed a special case of the updating task, the task of adding new information into an existing document. The information insertion task is formulated as a ranking problem and we applied some ranking models for the task. Moreover, we investigated processing methods for the information insertion task on two kinds of datasets: Wikipedia insertion data obtained from the previous work and Legal data built by ourselves. The experiment results show that proposed processing methods are also applicable for the new Legal dataset.

In Legal domain, there is no dataset for the information update task. Therefore, we built the Legal data from the data in the U.S Code in an automatic way. Despite the problems caused by the synthetic way of building the Legal dataset, it is good enough for the purpose of evaluating our proposed methods.

In Natural Language Processing, semantic relations between words can be exploited when measuring semantic text similarity of two text segments. In our research, we proposed a method of measuring topical overlap between two text segments, which incorporates word clusters, and used these similarity measures as additional semantic features in the learning models. Our method is somewhat similar to the method of semantic indexing with WordNet synsets [15, 25] in the idea of using intermediate word representations.

The advantage of our method is that it only used word clusters derived from unlabeled text data, without using any annotated lexical database like WordNet. We conducted experiments with various setting on both Wikipedia dataset, Legal dataset and reported results. The experiment results showed that combining cluster-based features and baseline features can boost the performance of the information insertion task on two datasets.

There are some problems with our method of using word clustering. First, the Brown word clustering algorithm generates a hard clustering in which each word in a word vocabulary is assigned to only one class. In fact, a word may have some different senses. Thus, strictly assigning a word to only one class may lead to the problem of word sense ambiguity and decrease the accuracy of measuring topical overlap between two text segments. As the future plan, we would like to apply soft word clustering algorithms such as [23].

The second problem is the coverage of the word clusters set. The text corpus on which we run the Word clustering algorithm is required to be large enough to cover our insertion data. In experiments, we obtained less improvement on the new Legal dataset with cluster-based features than on the Wikipedia dataset because of the out of domain problem. Many terms in Legal data do not appear in the vocabulary of the word clusters corpus that we used in experiments.

Finally, our method did not make use of word-to-word similarity metrics when measuring text-to-text similarity. The benefit of combining word-to-word similarity metrics into a text-to-text similarity metric was presented in [9] for some tasks. Although we cannot know in advance the effect of the method of using word-to-word similarity metrics for our task, it is still interesting to apply that method for our task.

Study Plan for Doctoral Course

In master's research, we did not consider the problem of information redundancy and information contradiction. Since new information may be redundant or contrastive with the original document, the task of recognizing information redundancy and information contradiction in text is very important to guarantee the consistency of updates. In the doctoral course, we would like to study the task of recognizing these phenomena in texts. We also consider the task of updating information in the general setting, dealing with other kinds of updating operations: deletion and modification.

Bibliography

[1] A. Baeza-Yates, R., Ribeiro-Neto B. (1999). Modern Information Retrieval, Addison-Wesley Longman Publishing Co., Inc., Boston, MA.

[2] Baker, L. D., McCallum, A. K. (1998). Distributional clustering of words for text classification. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 96–103.

[3] Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., Szpektor, I. (2006). The Second PASCAL Recognising Textual Entailment Challenge. In: Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venice, Italy.

[4] Bekkerman, R., El-Yaniv, R., Tishby, N., Winter, Y. (2003). Distributional word clusters vs. words for text categorization. The Journal of Machine Learning Research.

[5] Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n-gram models of natural language. Computational Linguistics, 18(4), pp. 467-479.

[6] Chen, E., Snyder, B., and Barzilay, R. (2007). Incremental text structuring with online hierachical ranking. In Proceedings of the EMNLP, pp. 83–91.

[7] Chen, E. (2008). Discourse Models for Collaboratively Edited Corpora. Masters thesis, Massachusetts Institute of Technology.

[8] Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In Proceedings of the EMNLP, pp. 1-8.

[9] Corley, C., Mihalcea, R. (2005). Measuring the Semantic Similarity of Texts. Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pages 13–18, Ann Arbor.

[10] Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y. (2006). Online Passive-Aggressive Algorithms. The Journal of Machine Learning Research, vol. 7, pp. 551–585.

[11] Dagan, I., Glickman, O., Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. In: Quionero-Candela, J., Dagan, I., Magnini, B., dAlch-Buc, F. (eds.) MLCW 2005. LNCS (LNAI), vol. 3944, pp. 177-190. Springer, Heidelberg.

[12] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, vol. 41, pp. 391–407.

[13] Freund, Y., and Schapire, R. E. (1999). Large margin classification using the perceptron algorithm. Machine Learning, 37(3):277-296.

[14] Giampiccolo, D., Magnini, B., Dagan, I., Dolan, B. (2007). The Third PASCAL Recognizing Textual Entailment Challenge. In: ACL-PASCAL Workshop on Textual Entailment and Paraphrasing.

[15] Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J. (1998). Indexing with wordnet synsets can improve text retrieval. In Proceedings of the COLING/ACL98 Workshop on Usage of WordNet for NLP.

[16] Jurafsky, D., and Martin, J. H. (2008). Speech and Language Processing. Prentice-Hall, New Jersey, USA.

[17] Katayama, T. (2007). Legal engineering an engineering approach to laws in e-society age. In: Proc. of the 1st Intl. Workshop on JURISIN.

[18] Katayama, T., Shimazu, A., Tojo, S., Futatsugi, K., Ochimizu, K. (2008). e-Society and legal engineering (in Japanese). Journal of the Japanese Society for Artificial Intelligence 23(4), 529-536.

[19] Kim, H. D., and Zhai, C. (2009). Generating Comparative Summaries of Contradictory Opinions in Text. In Proceedings of the 18th ACM International Conference on Information and Knowledge Management (CIKM'09), Hongkong, pp. 385–394.

[20] Kimura, Y., Nakamura, M., Shimazu, A. (2008). Treatment of legal sentences including itemized and referential expressions towards translation into logical forms. In: Proc. of the 2nd Intl. Workshop on JURISIN, pp. 73-82.

[21] Koo, T., Carreras, X., and Collins, M. (2008). Simple semi-supervised dependency parsing. In Proceedings of ACL-08, pp. 595-603.

[22] Lapata, M., and Barzilay, R. (2005). Automatic evaluation of text coherence: Models and representations. In IJCAI, pages 1085-1090.

[23] Li, W., and McCallum, A. (2005). Semi-supervised sequence modeling with syntactic topic models. In Proceedings of Twentieth National Conference on Artificial Intelligence, pp. 813–818.

[24] Liang, P. (2005). Semi-Supervised Learning for Natural Language. Masters thesis, Massachusetts Institute of Technology.

[25] Mihalcea, R., Moldovan, D. (2000). Semantic indexing using WordNet senses. In Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, October 08–08, 2000, Hong Kong.

[26] Miller, George A. WordNet - About Us. (1009). WordNet-Princeton University. Website: `http://wordnet.princeton.edu`

[27] Miller, S., Guinness, J., and Zamanian, A. (2008). Name tagging with word clusters and discriminative training. In Proceedings of HLT-NAACL, pp. 337-342.

[28] Nakamura, M., Nobuoka, S., Shimazu, A. (2008). Towards translation of legal sentences into logical forms. In Satoh, K., Inokuchi, A., Nagao, K., Kawamura, T., eds.: New Frontiers in Artificial Intelligence: JSAI 2007 Conference andWorkshops, Miyazaki, Japan, June 18-22, 2007, Revised Selected Papers. Volume 4914 of Lecture Notes in Computer Science., Springer, pp. 349-362.

[29] Nakamura, M., Kimura, Y., Pham, M. Q. N., Nguyen, M. L., and Shimazu, A. (2008). Treatment of Legal Sentences Including Itemization Written in Japanese, English and Vietnamese. In Proc. of the EMALP Workshop, PRICAI 2008, Hanoi, Vietnam, pp.102–113.

[30] Ogawa, Y., Inagaki, S., Toyama, K. (2008). Automatic Consolidation of Japanese Statutes Based on Formalization of Amendment Sentences. In: Satoh, K., Inokuchi, A., Nagao, K., Kawamura, T. (eds.) JSAI 2007. LNCS, vol. 4914, pp. 349-362. Springer, Heidelberg.

[31] Pham, M. Q. N., Nguyen, M. L., Shimazu, A. (2009). Incremental Text Structuring with Word Clusters. In Proceedings of the Conference of the Pacific Association for Computational Linguistics 2009, Hokkaido, Japan, pp. 109–114.

[32] Pham, M. Q. N., Nguyen, M. L., Shimazu, A. (2010). The Information Insertion Task with Intermediate Word Representation. The 16$^{\text{th}}$ NLP Annual Meeting, Tokyo, 2010, March. (to appear)

[33] Ponte J. M., Croft W. B. (1998). A language modeling approach to information retrieval. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275–281.

[34] Qiu, Y., Frei, H. (1993). Concept based query expansion. In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 160–169.

[35] Radev, D. R., Jing, H., and Budzikowska, M. (2000). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In ANLP/NAACL Workshop on Summarization Seattle, WA.

[36] Robertson, S., Zaragoza, H., Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. In: Proc. of the thirteenth ACM international conference on Information and knowledge management, pp. 42–49.

[37] Toutanova, K., and Manning, C. D. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63–70.

[38] Toutanova, K., Klein, D., Manning, C., and Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252–259.

[39] Preliminary Recommendations on Semantic Encoding Interim Report (1998). Retrieved January 2010 from the Website of Expert Advisory Group on Language Engineering Standards: `http://www.ilc.cnr.it/EAGLES96/rep2/node37.html`

[40] e-Government. Retrieved from the Wikipedia: `http://en.wikipedia.org/wiki/E_government`

[41] United States Code. Retrieved from the Website of the U.S. Government Printing Office: `http://www.gpoaccess.gov/uscode/about.html`

[42] Website of Office of the Law Revision Counsel. The United States Code. Retrieved from `http://uscode.house.gov/lawrevisioncounsel.shtml`.

[43] Website of Essem Educational. Text Coherence and Cohesion. Retrieved from `http://www.readability.biz/Coherence2.html`

Appendix A

Legal Dataset

Title No#	**Name**
Title 1	General Provisions
Title 3	The President
Title 4	Flag and Seal, Seat of Government, and the States
Title 5	Government Organization and Employees
Title 6	Domestic Security
Title 9	Arbitration
Title 11	Bankruptcy
Title 13	Census
Title 17	Conservation
Title 24	Highways
Title 27	Intoxicating Liquors
Title 32	National Guard
Title 35	Patents
Title 37	Pay and Allowances of the Uniformed Services
Title 39	Postal Service
Title 41	Public Contracts
Title 44	Public Printing and Documents

Table A.1: List of legal documents used in Legal dataset

Original place	**Sentence**
Title 3 Section 11	On the day thereafter they shall forward by registered mail two of such certificates and lists to the Archivist of the United States at the seat of government, one of which shall be held subject to the order of the President of the Senate.
Title 3 Section 15	But if the two Houses shall disagree in respect of the counting of such votes, then, and in that case, the votes of the electors whose appointment shall have been certified by the executive of the State, under the seal thereof, shall be counted.
Title 4 Section 7	By order of the President, the flag shall be flown at half-staff upon the death of principal figures of the United States Government and the Governor of a State, territory, or possession, as a mark of respect to their memory.
Title 4 Section 114	No State may impose an income tax on any retirement income of an individual who is not a resident or domiciliary of such State (as determined under the laws of such State).
Title 17 Section 1303	Protection for a design under this chapter shall be available notwithstanding the employment in the design of subject matter excluded from protection under section 1302 if the design is a substantial revision, adaptation, or rearrangement of such subject matter.
Title 32 Section 301	To be eligible for Federal recognition as an enlisted member of the National Guard, a person must have the qualifications prescribed by the Secretary concerned for the grade, branch, position, and type of unit or organization involved.
Title 14 Section 832	Members of the Auxiliary who incur physical injury or contract sickness or disease while performing any duty to which they have been assigned by competent Coast Guard authority shall be entitled to the same hospital treatment afforded members of the Coast Guard.
Title 35 Section 301	If the person explains in writing the pertinency and manner of applying such prior art to at least one claim of the patent, the citation of such prior art and the explanation thereof will become a part of the official file of the patent.

Table A.2: Some insertion sentences in Legal dataset

TITLE 4 - FLAG AND SEAL, SEAT OF GOVERNMENT, AND THE STATES
CHAPTER 1 - THE FLAG
Sec. 7. Position and manner of display

The flag, when carried in a procession with another flag or Flags, should be either on the marching right; that is, the flag's own right, or, if there is a line of other flags, in front of the center of that line.
(a) The flag should not be displayed on a float in a parade except from a staff, or as provided in subsection (i) of this section.

(b) The flag should not be draped over the hood, top, sides, or back of a vehicle or of a railroad train or a boat. When the flag is displayed on a motorcar, the staff shall be fixed firmly to the chassis or clamped to the right fender.

(c) No other flag or pennant should be placed above or, if on the same level, to the right of the flag of the United States of America, except during church services conducted by naval chaplains at sea, when the church pennant may be flown above the flag during church services for the personnel of the Navy. No person shall display the flag of the United Nations or any other national or international flag equal, above, or in a position of superior prominence or honor to, or in place of, the flag of the United States at any place within the United States or any Territory or possession thereof: Provided, That nothing in this section shall make unlawful the continuance of the practice heretofore followed of displaying the flag of the United Nations in a position of superior prominence or honor, and other national flags in positions of equal prominence or honor, with that of the flag of the United States at the headquarters of the United Nations.

(d) The flag of the United States of America, when it is displayed with another flag against a wall from crossed staffs, should be on the right, the flag's own right, and its staff should be in front of the staff of the other flag.

(e) The flag of the United States of America should be at the center and at the highest point of the group when a number of flags of States or localities or pennants of societies are grouped and displayed from staffs.

(f) When flags of States, cities, or localities, or pennants of societies are flown on the same halyard with the flag of the United States, the latter should always be at the peak. When the flags are flown from adjacent staffs, the flag of the United States should be hoisted first and lowered last. No such flag or pennant may be placed above the flag of the United States or to the United States flag's right.

(g) When flags of two or more nations are displayed, they are to be flown from separate staffs of the same height. The flags should be of approximately equal size. International usage forbids the display of the flag of one nation above that of another nation in time of peace.

(h) When the flag of the United States is displayed from a staff projecting horizontally or at an angle from the window sill, balcony, or front of a building, the union of the flag should be placed at the peak of the staff unless the flag is at half-staff. When the flag is suspended over a sidewalk from a rope extending from a house to a pole at the edge of the sidewalk, the flag should be hoisted out, union first, from the building.

(i) When displayed either horizontally or vertically against a wall, the union should be uppermost and to the flag's own right, that is, to the observer's left. When displayed in a window, the flag should be displayed in the same way, with the union or blue field to the left of the observer in the street.

Figure A.1: An example insertion in Title 4 - Flag and Seal, Seat of Government, and the States (Part 1). An insertion example is shown in bold-face underlined font. The sentence is extracted from Section 7 - Position and manner of display.

(j) When the flag is displayed over the middle of the street, it should be suspended vertically with the union to the north in an east and west street or to the east in a north and south street.

(k) When used on a speaker's platform, the flag, if displayed flat, should be displayed above and behind the speaker. When displayed from a staff in a church or public auditorium, the flag of the United States of America should hold the position of superior prominence, in advance of the audience, and in the position of honor at the clergyman's or speaker's right as he faces the audience. Any other flag so displayed should be placed on the left of the clergyman or speaker or to the right of the audience.
(l) The flag should form a distinctive feature of the ceremony of unveiling a statue or monument, but it should never be used as the covering for the statue or monument.

(m) The flag, when flown at half-staff, should be first hoisted to the peak for an instant and then lowered to the half-staff position. The flag should be again raised to the peak before it is lowered for the day. On Memorial Day the flag should be displayed at half-staff until noon only, then raised to the top of the staff. **By order of the President, the flag shall be flown at half-staff upon the death of principal figures of the United States Government and the Governor of a State, territory, or possession, as a mark of respect to their memory**. In the event of the death of other officials or foreign dignitaries, the flag is to be displayed at half-staff according to Presidential instructions or orders, or in accordance with recognized customs or practices not inconsistent with law. In the event of the death of a present or former official of the government of any State, territory, or possession of the United States or the death of a member of the Armed Forces from any State, territory, or possession who dies while serving on active duty, the Governor of that State, territory, or possession may proclaim that the National flag shall be flown at half-staff, and the same authority is provided to the Mayor of the District of Columbia with respect to present or former officials of the District of Columbia and members of the Armed Forces from the District of Columbia. When the Governor of a State, territory, or possession, or the Mayor of the District of Columbia, issues a proclamation under the preceding sentence that the National flag be flown at half-staff in that State, territory, or possession or in the District of Columbia because of the death of a member of the Armed Forces, the National flag flown at any Federal installation or facility in the area covered by that proclamation shall be flown at half-staff consistent with that proclamation. The flag shall be flown at half-staff 30 days from the death of the President or a former President; 10 days from the day of death of the Vice President, the Chief Justice or a retired Chief Justice of the United States, or the Speaker of the House of Representatives; from the day of death until interment of an Associate Justice of the Supreme Court, a Secretary of an executive or military department, a former Vice President, or the Governor of a State, territory, or possession; and on the day of death and the following day for a Member of Congress. The flag shall be flown at half-staff on Peace Officers Memorial Day, unless that day is also Armed Forces Day. As used in this subsection -

(1) the term "half-staff" means the position of the flag when it is one-half the distance between the top and bottom of the staff;
(2) the term "executive or military department" means any agency listed under sections 101 and 102 of title 5, United States Code; and
(3) the term "Member of Congress" means a Senator, a Representative, a Delegate, or the Resident Commissioner from Puerto Rico.

(n) When the flag is used to cover a casket, it should be so placed that the union is at the head and over the left shoulder. The flag should not be lowered into the grave or allowed to touch the ground.

(o) When the flag is suspended across a corridor or lobby in a building with only one main entrance, it should be suspended vertically with the union of the flag to the observer's left upon entering. If the building has more than one main entrance, the flag should be suspended vertically near the center of the corridor or lobby with the union to the north, when entrances are to the east and west or to the east when entrances are to the north and south. If there are entrances in more than two directions, the union should be to the east.

Figure A.2: An example insertion in Title 4 - Flag and Seal, Seat of Government, and the States (Part 2). An insertion example is shown in bold-face underlined font. The sentence is extracted from Section 7 - Position and manner of display.

TITLE 14 - COAST GUARD
PART II - COAST GUARD RESERVE AND AUXILIARY
CHAPTER 23 - COAST GUARD AUXILIARY
Sec. 832. Injury or death in line of duty

When any member of the Auxiliary is physically injured or dies as a result of physical injury incurred while performing any duty to which he has been assigned by competent Coast Guard authority, such member or his beneficiary shall be entitled to the same benefits provided for temporary members of the Reserve who suffer physical injury or death resulting from physical injury incurred incident to service. **Members of the Auxiliary who incur physical injury or contract sickness or disease while performing any duty to which they have been assigned by competent Coast Guard authority shall be entitled to the same hospital treatment afforded members of the Coast Guard**. The performance of a duty as the term is used in this section includes time engaged in traveling back and forth between the place of assigned duty and the permanent residence of a member of the Auxiliary.

Figure A.3: An example insertion in Title 14 - Coast Guard. An insertion example is shown in bold-face underlined font. The sentence is extracted from Section 832 - Injury or death in line of duty.

TITLE 35 – PATENTS
PART II - PATENTABILITY OF INVENTIONS AND GRANT OF PATENTS
CHAPTER 17 - SECRECY OF CERTAIN INVENTIONS AND FILING APPLICATIONS IN FOREIGN COUNTRY
Sec. 184. Filing of application in foreign country

Except when authorized by a license obtained from the Commissioner of Patents a person shall not file or cause or authorize to be filed in any foreign country prior to six months after filing in the United States an application for patent or for the registration of a utility model, industrial design, or model in respect of an invention made in this country. A license shall not be granted with respect to an invention subject to an order issued by the Commissioner of Patents pursuant to section 181 of this title without the concurrence of the head of the departments and the chief officers of the agencies who caused the order to be issued. The license may be granted retroactively where an application has been filed abroad through error and without deceptive intent and the application does not disclose an invention within the scope of section 181 of this title.

The term "application" when used in this chapter includes applications and any modifications, amendments, or supplements thereto, or divisions thereof.

The scope of a license shall permit subsequent modifications, amendments, and supplements containing additional subject matter if the application upon which the request for the license is based is not, or was not, required to be made available for inspection under section 181 of this title and if such modifications, amendments, and supplements do not change the general nature of the invention in a manner which would require such application to be made available for inspection under such section 181. In any case in which a license is not, or was not, required in order to file an application in any foreign country, such subsequent modifications, amendments, and supplements may be made, without a license, to the application filed in the foreign country if the United States application was not required to be made available for inspection under section 181 and if such modifications, amendments, and supplements do not, or did not, change the general nature of the invention in a manner which would require the United States application to have been made available for inspection under such section 181.

Figure A.4: An example insertion in Title 35 - Patents. An insertion example is shown in bold-face underlined font. The sentence is extracted from Section 184 - Filing of application in foreign country.

TITLE 37 - PAY AND ALLOWANCES OF THE UNIFORMED SERVICES
CHAPTER 10 - PAYMENTS TO MISSING PERSONS
Sec. 556. Secretarial determinations

(a) The Secretary concerned, or his designee, may make any determination necessary to administer this chapter and, when so made, it is conclusive as to -
(1) death or finding of death;
(2) the fact of dependency under this chapter;
(3) the fact of dependency for the purpose of paying six months' death gratuities authorized by law;
(4) the fact of dependency under any other law authorizing the payment of pay, allowances, or other emoluments to enlisted members of the armed forces, when the payments are contingent on dependency;
(5) any other status covered by this chapter;
(6) an essential date, including one on which evidence or information is received by the Secretary concerned; and
(7) whether information received concerning a member of a uniformed service is to be construed and acted on as an official report of death.

Paragraphs (1), (5), (6), and (7) only apply with respect to a case to which section 555 of this title applies.

(b) When the Secretary concerned, in a case to which section 555 of this title applies, receives information that he considers establishes conclusively the death of a member of a uniformed service, he shall, notwithstanding any earlier action relating to death or other status of the member, act on it as an official report of death. After the end of the 12-month period in a missing status prescribed by section 555 of this title, the Secretary concerned, or his designee, shall, when he considers that the information received, or a lapse of time without information, establishes a reasonable presumption that a member in a missing status is dead, make a finding of death.

(c) The Secretary concerned, or his designee, may determine the entitlement of a member to pay and allowances under this chapter, including credits and charges in his account, and that determination is conclusive. **An account may not be charged or debited with an amount that a member captured, beleaguered, or besieged by a hostile force may receive or be entitled to receive from, or have placed to his credit by, the hostile force as pay, allowances, or other compensation.**

(d) The Secretary concerned, or his designee, may, when warranted by the circumstances, reconsider a determination made under this chapter, and change or modify it.

(e) When the account of a member has been charged or debited with an allotment paid under this chapter, the amount so charged or debited shall be recredited to the account of the member if the Secretary concerned, or his designee, determines that the payment was induced by fraud or misrepresentation to which the member was not a party.

(f) Except an allotment for an unearned insurance premium, an allotment paid from pay and allowances of a member for the period he is entitled to pay and allowances under section 552 of this title may not be collected from the allottee as an overpayment when it was caused by delay in receiving evidence of death. An allotment payment for a period after the end of entitlement to pay and allowances under this chapter, or otherwise, which was caused by delay in receiving evidence of death, may not be collected from the allottee or charged against the pay of the deceased member.

(g) The Secretary concerned, or his designee, may waive the recovery of an erroneous payment or overpayment of an allotment to a dependent if he considers recovery is against equity and good conscience.

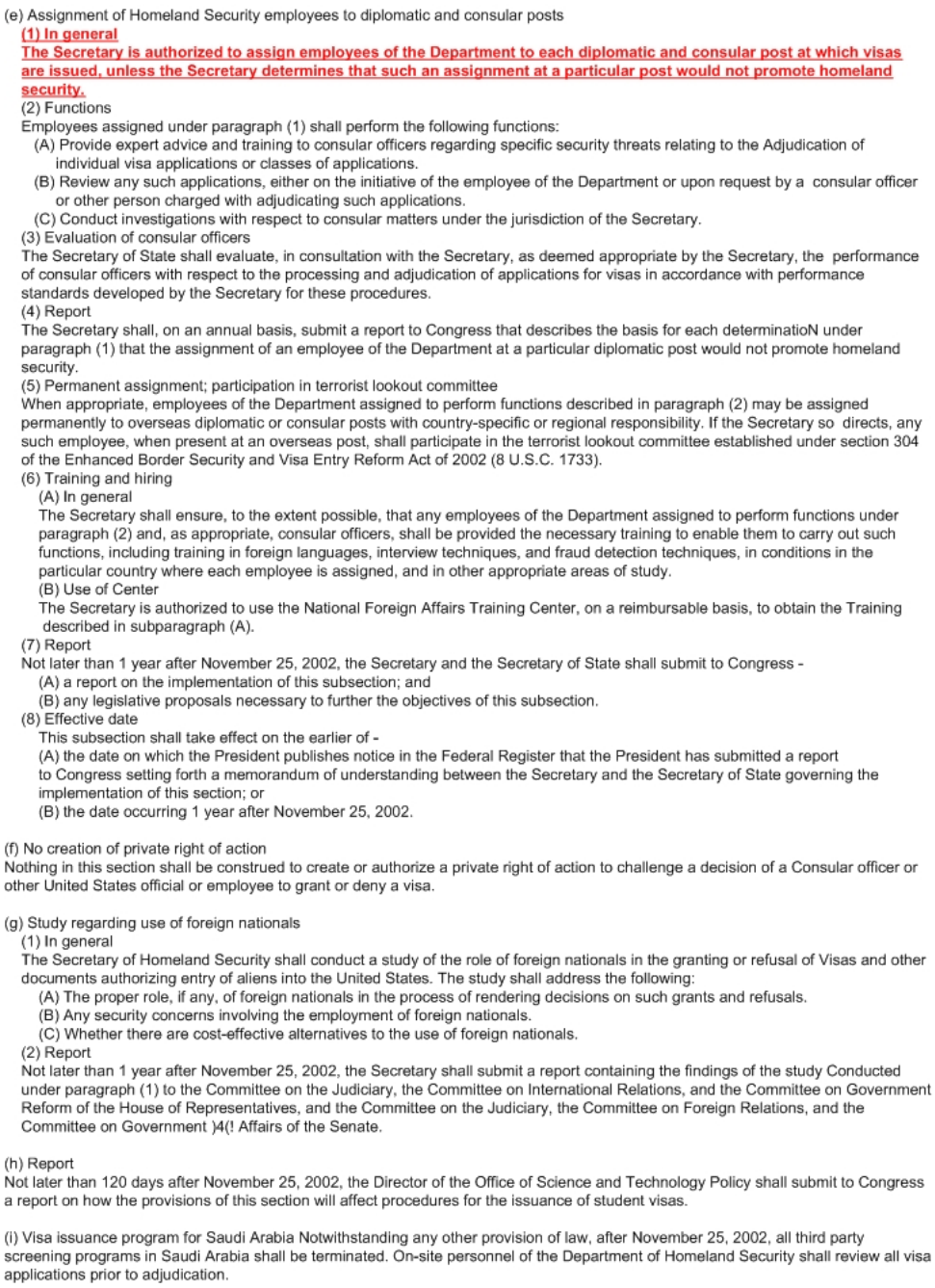
(h) For the sole purpose of determining pay under this section, a dependent of a member of a uniformed service on active duty is treated as if he were a member. Any determination made by the Secretary concerned, or his designee, under this section in a case to which section 555 of this title applies is conclusive on all other departments and agencies of the United States. This subsection does not entitle a dependent to pay, allowances, or other compensation to which he is not otherwise entitled.

Figure A.5: An example insertion in Title 37 - Pay and allowances of the uniformed services. An insertion example is shown in bold-face underlined font. The sentence is extracted from Section 556 - Secretarial determinations.

TITLE 6 - DOMESTIC SECURITY
CHAPTER 1 - HOMELAND SECURITY ORGANIZATION
SUBCHAPTER IV - DIRECTORATE OF BORDER AND TRANSPORTATION SECURITY
Part C - Miscellaneous Provisions
Sec. 236. Visa issuance

(a) Definition
In this subsection,)1(! The term "consular office") 2(! has the meaning given that term under section 101(a)(9) of the Immigration and Nationality Act (8 U.S.C. 1101(a)(9)).
(b) In general
Notwithstanding section 104(a) of the Immigration and Nationality AcT (8 U.S.C. 1104(a)) or any other provision of law, and except as provided in subsection © of this section, the Secretary -
(1) shall be vested exclusively with all authorities to issue regulations with respect to, administer, and enforce the provisions of such Act [8 U.S.C. 1101 et seq.], and of all other immigration and nationality laws, relating to the functions of consular officers of the United States in connection with the granting or refusal of visas, and shall have the authority to refuse visas in accordance with law and to develop programs of homeland security training for consular officers (in addition to consular training provided by the Secretary of State), which authorities shall be exercised through the Secretary of State, except that the Secretary shall not have authority to alter or reverse the decision of a consular officer to refuse a visa to an alien; and
(2) shall have authority to confer or impose upon any officer or employee of the United States, with the consent of the head of the executive agency under whose jurisdiction such officer or employee is serving, any of the functions specified in paragraph(1).
(c) Authority of the Secretary of State
(1) In general
Notwithstanding subsection (b) of this section, the Secretary of State may direct a consular officer to refuse a visa to an alien if the Secretary of State deems such refusal necessary or advisable in the foreign policy or security interests of the United States.
(2) Construction regarding authority
Nothing in this section, consistent with the Secretary of Homeland Security's authority to refuse visas in accordance with law, shall be construed as affecting the authorities of the Secretary of State under the following provisions of law:
(A) Section 101(a)(15)(A) of the Immigration and Nationality Act (8 U.S.C. 1101(a)(15)(A)).
(B) Section 204(d)(2) of the Immigration and Nationality Act (8 U.S.C. 1154) (as it will take effect upon the entry into force of the Convention on Protection of Children and Cooperation in Respect to Inter-Country adoption
(C) Section 212(a)(3)(B)(i)(IV)(bb) of the Immigration and Nationality Act (8 U.S.C. 1182(a)(3)(B)(i)(IV)(bb)).
(D) Section 212(a)(3)(B)(i)(VI) of the Immigration and Nationality Act (8 U.S.C. 1182(a)(3)(B)(i)(VI)).
(E) Section 212(a)(3)(B)(vi)(II) of the Immigration and Nationality Act (8 U.S.C. 1182(a)(3)(B)(vi)(II)).
(F) Section 212(a)(3)(C) of the Immigration and Nationality Act (8 U.S.C. 1182(a)(3)(C)).
(G) Section 212(a)(10)(C) of the Immigration and Nationality Act (8 U.S.C. 1182(a)(10)(C)).
(H) Section 212(f) of the Immigration and Nationality Act (8 U.S.C. 1182(f)).
(I) Section 219(a) of the Immigration and Nationality Act (8 U.S.C. 1189(a)).
(J) Section 237(a)(4)(C) of the Immigration and Nationality Act (8 U.S.C. 1227(a)(4)(C)).
(K) Section 401 of the Cuban Liberty and Democratic Solidarity (LIBERTAD) Act of 1996 [22 U.S.C. 6091].
(L) Section 613 of the Departments of Commerce, Justice, and State, the Judiciary and Related Agencies Appropriations Act, 1999)3(! (as contained in section 101(b) of division A of Public Law 105-277) (Omnibus Consolidated and Emergency Supplemental Appropriations Act, 1999); 112 Stat. 2681; H.R.4328 (originally H.R. 4276) as amended by section 617 of Public Law 106-553.
(M) Section 103(f) of the Chemical Weapon Convention Implementation Act of 1998 [22 U.S.C. 6713(f)] (112 Stat. 2681-865).
(N) Section 801 of H.R. 3427, the Admiral James W. Nance and Meg Donovan Foreign Relations Authorization Act, Fiscal Years 2000 and 2001 [8 U.S.C. 1182e], as enacted by reference in Public Law 106-113.
(O) Section 568 of the Foreign Operations, Export Financing, and Related Programs Appropriations Act, 2002 (Public Law 107-115).
(P) Section 51 of the State Department Basic Authorities Act of 1956 (22 U.S.C. 2723).
(d) Consular officers and chiefs of missions
(1) In general
Nothing in this section may be construed to alter or affect -
(A) the employment status of consular officers as employees of the Department of State; or
(B) the authority of a chief of mission under section 207 of the Foreign Service Act of 1980 (22 U.S.C. 3927).
(2) Construction regarding delegation of authority
Nothing in this section shall be construed to affect any delegation of authority to the Secretary of State by the President pursuant to any proclamation issued under section 212(f) of the Immigration and Nationality Act (8 U.S.C.1182(f)), consistent with the Secretary of Homeland Security's authority to refuse visas in accordance with law.

Figure A.6: An example insertion in Title 6 - Domestic Security (Part 1). An insertion example is shown in bold-face underlined font. The sentence is extracted from Section 236 - Visa issuance.



(e) Assignment of Homeland Security employees to diplomatic and consular posts

(1) In general

The Secretary is authorized to assign employees of the Department to each diplomatic and consular post at which visas are issued, unless the Secretary determines that such an assignment at a particular post would not promote homeland security.

(2) Functions

Employees assigned under paragraph (1) shall perform the following functions:

(A) Provide expert advice and training to consular officers regarding specific security threats relating to the Adjudication of individual visa applications or classes of applications.

(B) Review any such applications, either on the initiative of the employee of the Department or upon request by a consular officer or other person charged with adjudicating such applications.

(C) Conduct investigations with respect to consular matters under the jurisdiction of the Secretary.

(3) Evaluation of consular officers

The Secretary of State shall evaluate, in consultation with the Secretary, as deemed appropriate by the Secretary, the performance of consular officers with respect to the processing and adjudication of applications for visas in accordance with performance standards developed by the Secretary for these procedures.

(4) Report

The Secretary shall, on an annual basis, submit a report to Congress that describes the basis for each determinatioN under paragraph (1) that the assignment of an employee of the Department at a particular diplomatic post would not promote homeland security.

(5) Permanent assignment; participation in terrorist lookout committee

When appropriate, employees of the Department assigned to perform functions described in paragraph (2) may be assigned permanently to overseas diplomatic or consular posts with country-specific or regional responsibility. If the Secretary so directs, any such employee, when present at an overseas post, shall participate in the terrorist lookout committee established under section 304 of the Enhanced Border Security and Visa Entry Reform Act of 2002 (8 U.S.C. 1733).

(6) Training and hiring

(A) In general

The Secretary shall ensure, to the extent possible, that any employees of the Department assigned to perform functions under paragraph (2) and, as appropriate, consular officers, shall be provided the necessary training to enable them to carry out such functions, including training in foreign languages, interview techniques, and fraud detection techniques, in conditions in the particular country where each employee is assigned, and in other appropriate areas of study.

(B) Use of Center

The Secretary is authorized to use the National Foreign Affairs Training Center, on a reimbursable basis, to obtain the Training described in subparagraph (A).

(7) Report

Not later than 1 year after November 25, 2002, the Secretary and the Secretary of State shall submit to Congress -

(A) a report on the implementation of this subsection; and

(B) any legislative proposals necessary to further the objectives of this subsection.

(8) Effective date

This subsection shall take effect on the earlier of -

(A) the date on which the President publishes notice in the Federal Register that the President has submitted a report to Congress setting forth a memorandum of understanding between the Secretary and the Secretary of State governing the implementation of this section; or

(B) the date occurring 1 year after November 25, 2002.

(f) No creation of private right of action

Nothing in this section shall be construed to create or authorize a private right of action to challenge a decision of a Consular officer or other United States official or employee to grant or deny a visa.

(g) Study regarding use of foreign nationals

(1) In general

The Secretary of Homeland Security shall conduct a study of the role of foreign nationals in the granting or refusal of Visas and other documents authorizing entry of aliens into the United States. The study shall address the following:

(A) The proper role, if any, of foreign nationals in the process of rendering decisions on such grants and refusals.

(B) Any security concerns involving the employment of foreign nationals.

(C) Whether there are cost-effective alternatives to the use of foreign nationals.

(2) Report

Not later than 1 year after November 25, 2002, the Secretary shall submit a report containing the findings of the study Conducted under paragraph (1) to the Committee on the Judiciary, the Committee on International Relations, and the Committee on Government Reform of the House of Representatives, and the Committee on the Judiciary, the Committee on Foreign Relations, and the Committee on Government)4(! Affairs of the Senate.

(h) Report

Not later than 120 days after November 25, 2002, the Director of the Office of Science and Technology Policy shall submit to Congress a report on how the provisions of this section will affect procedures for the issuance of student visas.

(i) Visa issuance program for Saudi Arabia Notwithstanding any other provision of law, after November 25, 2002, all third party screening programs in Saudi Arabia shall be terminated. On-site personnel of the Department of Homeland Security shall review all visa applications prior to adjudication.

Figure A.7: An example insertion in Title 6 - Domestic Security (Part 2). An insertion example is shown in bold-face underlined font. The sentence is extracted from Section 236 - Visa issuance.

TITLE 37 - PAY AND ALLOWANCES OF THE UNIFORMED SERVICES
CHAPTER 7 - ALLOWANCES
Sec. 404a. Travel and transportation allowances: temporary lodging expenses

(a) Payment or Reimbursement of Subsistence Expenses. - (1) Under regulations prescribed by the Secretaries concerned, a member of a uniformed service who is ordered to make a change of permanent station described in paragraph (2) shall be paid or reimbursed for subsistence expenses of the member and the member's dependents for the period (subject to subsection (c)) for which the member and dependents occupy temporary quarters incident to that change of permanent station.
(2) Paragraph (1) applies to the following:
(A) A permanent change of station from any duty station to a duty station in the United States (other than Hawaii or Alaska).
(B) A permanent change of station from a duty station in the United States (other than Hawaii or Alaska) to a duty station outside the United States or in Hawaii or Alaska.
(C) In the case of a member who is reporting to the member's first permanent duty station, the change from the member's home of record or initial technical school to that first permanent duty station.

(b) Payment in Advance. - The Secretary concerned may make any payment for subsistence expenses to a member under this section in advance of the member actually incurring the expenses. **The amount of an advance payment made to a member shall be computed on the basis of the Secretary's determination of the average number of days that members and their dependents occupy temporary quarters under the circumstances applicable to the member and the member's dependents.**

(c) Maximum Payment Period. - (1) In the case of a change of permanent station described in subparagraph (A) or (C) of subsection (a)(2), the period for which subsistence expenses are to be paid or reimbursed under this section may not exceed 10 days.
(2) In the case of a change of permanent station described in subsection (a)(2)(B) -
(A) the period for which such expenses are to be paid or reimbursed under this section may not exceed five days; and
(B) such payment or reimbursement may be provided only for expenses incurred before leaving the United States (other than Hawaii or Alaska).
(3) Whenever the conditions described in clause (i) or (ii) of subparagraph (A) of section 403(b)(7) of this title exist for a military housing area or portion thereof, the Secretary concerned may increase the period for which subsistence expenses are to be paid or reimbursed under this section in the case of a change of permanent station described in subparagraph (A) or (C) of subsection (a)(2) in the same military housing area or portion thereof to a maximum of 20 days.

(d) Daily Subsistence Rates. - Regulations prescribed under subsection (a) shall prescribe average daily subsistence rates for purposes of this section for the member and for each dependent. Such rates may not exceed the maximum per diem rates prescribed under section 404(d) of this title for the area where the temporary quarters are located.

(e) Maximum Daily Payment. - A member may not be paid or reimbursed more than $180 a day under this section.

Figure A.8: An example insertion in Title 37 - Pay and allowances of the uniformed services. An insertion example is shown in bold-face underlined font. The sentence is extracted from Section 404a - Travel and transportation allowances: temporary lodging expenses.