International Journal of

Computational Linguistics & Chinese Language Processing

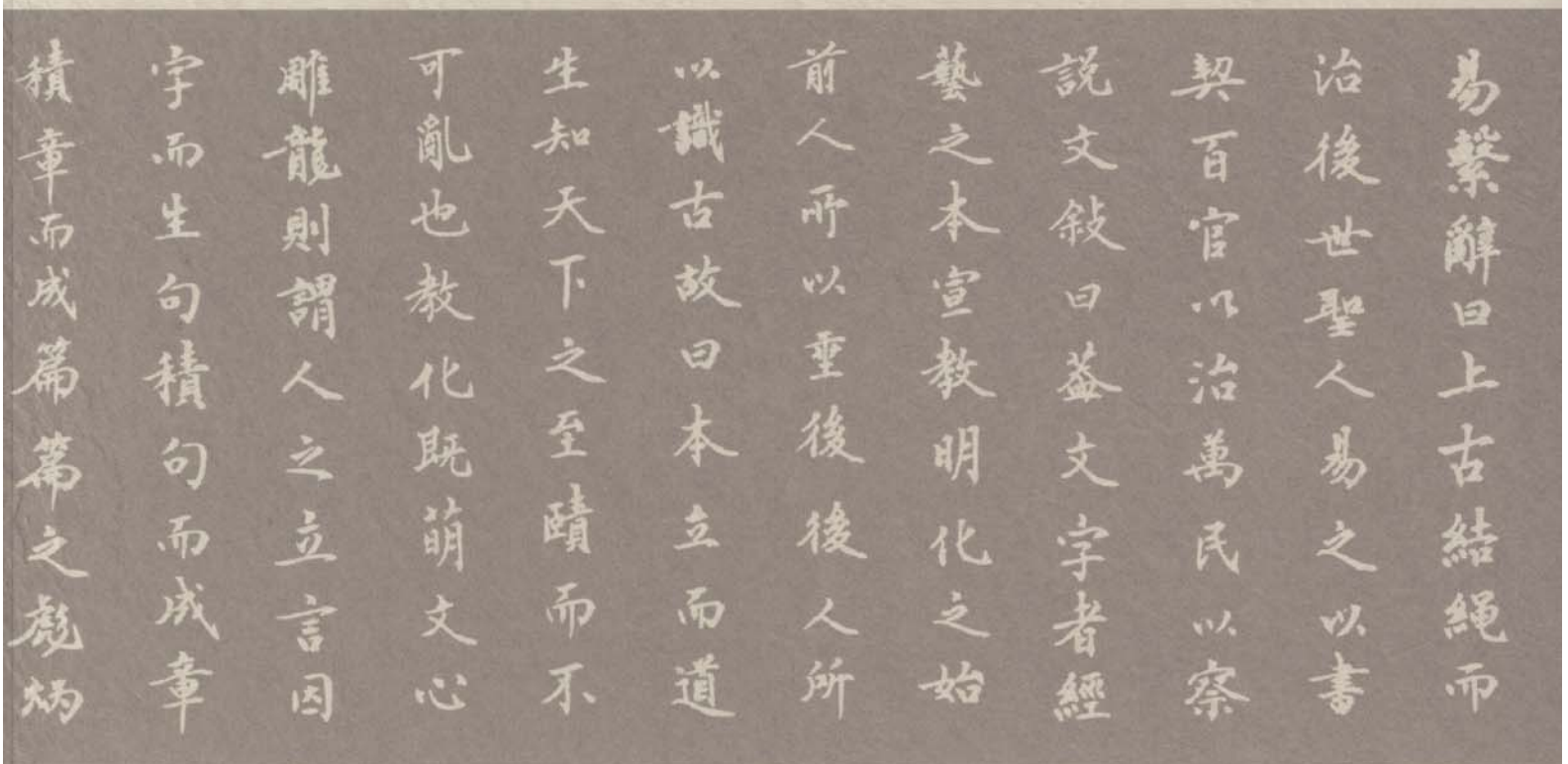
中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing

This journal is included in THCI, Linguistics Abstracts, and ACL Anthology.

Special Issue on "Selected Papers from ROCLING XXIV"

Guest Editors: Liang-Chih Yu, Richard Tzong-Han Tsai, Chia-Ping Chen, Cheng-Zen Yang and Shu-Kai Hsieh



Vol.17 No.4 December 2012 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

The Association for Computational Linguistics and Chinese Language Processing, Taipei

International Journal of

Computational Linguistics & Chinese Language Processing

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is an international journal published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). This journal was founded in August 1996 and is published four issues per year since 2005. This journal covers all aspects related to computational linguistics and speech/text processing of all natural languages. Possible topics for manuscript submitted to the journal include, but are not limited to:

- Computational Linguistics
- Natural Language Processing
- Machine Translation
- Language Generation
- Language Learning
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Information Retrieval and Extraction
- Web Information Extraction/Mining
- Corpus Linguistics
- Multilingual/Cross-lingual Language Processing

Membership & Subscriptions

If you are interested in joining ACLCLP, please see appendix for further information.

Copyright

Cover

Calligraphy by Professor Ching-Chun Hsieh, founding president of ACLCLP
Text excerpted and compiled from ancient Chinese classics, dating back to 700 B.C.
This calligraphy honors the interaction and influence between text and language

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 17, no. 4, December 2012

Contents

Special Issue Articles:
Selected Papers from ROCLING XXIV

Papers

Forewords

The 24th Conference on Computational Linguistics and Speech Processing (ROCLING 2012) was held at Yuan Ze University, on September 21-22, 2012. ROCLING is the leading and most comprehensive conference on computational linguistics and speech processing in Taiwan, bringing together researchers, scientists and industry participants to present their work and discuss recent trends in the field. This special issue presents extended and reviewed versions of five papers meticulously selected from ROCLING 2012.

The first paper "Detecting and Correcting Syntactic Errors in Machine Translation Using Feature-Based Lexicalized Tree Adjoining Grammars" proposes the use of *tree adjoining grammars* (TAG) to simultaneously detect and correct multiple ungrammatical types for machine translation. The second paper "TQDL: Integrated Models for Cross-Language Document Retrieval" proposes a framework integrating four statistical methods: **T**ranslation model, **Q**uery generation model, **D**ocument retrieval model and **L**ength Filter model for cross-language information retrieval (CLIR). The four independent methods work together to deal with the term disambiguation, query generation and document retrieval. The third paper "Domain Dependent Word Polarity Analysis for Sentiment Classification" explores the polarity of words in the corpora from three different application domains: real estate, hotel and restaurant, and then proposes a method to capture their sentiment differences. The fourth paper "Exploiting Machine Learning Models for Chinese Legal Documents Labeling, Case Classification, and Sentencing Prediction" discusses various interesting topics for Chinese legal document processing such as robbery and intimidation case classification. The fifth paper "Speech Recognition Leveraging Histogram Equalization Methods" uses a histogram equalization (HEQ) method for speech feature normalization to reduce the word error rate in speech recognition.

The Guest Editors of this special issue would like to thank all of the authors and reviewers for their contributions. We would also like to thank all the researchers and participants for sharing their knowledge and experience at the conference.

Guest Editors
Liang-Chih Yu
Department of Information Management, Yuan Ze University, Taiwan, R.O.C.
Richard Tzong-Han Tsai
Department of Computer Science and Engineering, Yuan Ze University, Taiwan, R.O.C.
Chia-Ping Chen
Department of Computer Science and Engineering, National Sun Yat-Sen University, Taiwan, R.O.C.
Cheng-Zen Yang
Department of Computer Science and Engineering, Yuan Ze University, Taiwan, R.O.C.
Shu-Kai Hsieh
Graduate Institute of Linguistics, National Taiwan University, Taiwan, R.O.C.

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 4, December 2012, pp. 1-14

Detecting and Correcting Syntactic Errors in Machine Translation Using Feature-Based Lexicalized Tree Adjoining Grammars

Wei-Yun Ma*, and Kathleen McKeown*

Abstract

Statistical machine translation has made tremendous progress over the past ten years. The output of even the best systems, however, is often ungrammatical because of the lack of sufficient linguistic knowledge. Even when systems incorporate syntax in the translation process, syntactic errors still result. To address this issue, we present a novel approach for detecting and correcting ungrammatical translations. In order to simultaneously detect multiple errors and their corresponding words in a formal framework, we use feature-based lexicalized tree adjoining grammars, where each lexical item is associated with a syntactic elementary tree, in which each node is associated with a set of feature-value pairs to define the lexical item's syntactic usage. Our syntactic error detection works by checking the feature values of all lexical items within a sentence using a unification framework. In order to simultaneously detect multiple error types and track their corresponding words, we propose a new unification method which allows the unification procedure to continue when unification fails and also to propagate the failure information to relevant words. Once error types and their corresponding words are detected, one is able to correct errors based on a unified consideration of all related words under the same error types. In this paper, we present some simple mechanism to handle part of the detected situations. We use our approach to detect and correct translations of six single statistical machine translation systems. The results show that most of the corrected translations are improved.

Keywords: Machine Translation, Syntactic Error, Post Editing, Tree Adjoining Grammar, Unification.

* Department of Computer Science, Columbia University, New York, USA
E-mail: {ma, kathy}@cs.columbia.edu

1. Introduction

Statistical machine translation has made tremendous progress over the past ten years. The output of even the best systems, however, is often ungrammatical because of the lack of sufficient linguistic knowledge. Even when systems incorporate syntax in the translation process, syntactic errors still result. We have developed a novel, post-editing approach which features: 1) the use of XTAG grammar, a rule-based grammar developed by linguists, 2) the ability to simultaneously detect multiple ungrammatical types and their corresponding words by using unification of feature structures, and 3) the ability to simultaneously correct multiple ungrammatical types based on the detection information. To date, we have developed the infrastructure for this approach and demonstrated its utility for agreement errors.

As illustrative examples, consider the following three ungrammatical English sentences:

1. Many young student play basketball.
2. John play basketball and Tom also play basketball.
3. John thinks to play basketball.

In 1 and 2 above, number agreement errors between the subjects and verbs (and quantifier) cause the sentences to be ungrammatical, while in 3, the infinitive following the main verb makes it ungrammatical. One could argue that an existing grammar checker could do the error detection for us, but if we use Microsoft Word 2010 (MS Word)'s grammar checker (Heidorn, 2000) to check the three sentences, the entire first sentence will be underlined with green wavy lines without any indication of what should be corrected, while no errors are detected in 2 and 3.

The grammar we use is based on a feature-based lexicalized tree adjoining grammars (FB-LTAG) English grammar, named XTAG grammar (XTAG group, 2001). In FB-LTAG, each lexical item is associated with a syntactic elementary tree, in which each node is associated with a set of feature-value pairs, called Attribute Value Matrices (AVMs). AVMs define the lexical item's syntactic usage. Our syntactic error detection works by checking the AVM values of all lexical items within a sentence using a unification framework. Thus, we use the feature structures in the AVMs to detect the error type and corresponding words. In order to simultaneously detect multiple error types and track their corresponding words, we propose a new unification method which allows the unification procedure to continue when unification fails and also to propagate the failure information to relevant words. We call the modified unification a *fail propagation unification.*

2. Related Work

Grammar checking is mostly used in word processors as a writing aid. Three methods are widely used for grammar checking given a sentence: statistic-based checking, rule-based checking and syntax-based checking. In statistic-based checking, POS tag sequences (Atwell & Elliot, 1987) or an N-gram language model (Alam *et al*., 2006; Wu *et al*., 2006) is trained from a training corpus and uncommon sequences in the training corpus are considered incorrect. Huang *et al*. (2010) extracted erroneous and correct patterns of consecutive words from the data of an online-editing diary website. In rule-based checking, a set of hand crafted rules out of words, POS tags and chucks (Naber, 2003) or parsing results (Heidorn, 2000) are designed to detect errors. In syntax-based checking, Jensen *et al*. (1993) utilize a parsing procedure to detect errors: each sentence must be syntactically parsed; a sentence is considered incorrect if parsing does not succeed.

Focusing on machine translation's grammar checking, Stymne and Ahrenberg (2010) utilized an existing rule-based Swedish grammar checker, as a post-processing tool for their English-Swedish translation system. They tried to fix the ungrammatical translation parts by applying the grammar checker's correction suggestions. In contrast of their using an existing grammar checker, we developed our own novel grammar checker for translated English in order to better controlling the quality of error detection, error types, and the directions of error correction in translation context.

Our approach is a mix of rule-based checking and syntax-based checking: The XTAG English grammar is designed by linguists while the detecting procedure is based on syntactic operations which dynamically reference the grammar. The work could be regarded as an extension of (Ma & McKeown, 2011), in which grammatical error detection based on XTAG English grammar is carried out to filter out ungrammatical combined translations in their framework of system combination for machine translation. In contrast of (Ma & McKeown, 2011), our approach is not only capable to detect grammatical errors, but also has the capability of identifying error types and errors' causes, and correcting certain cases of errors.

3. Background

We briefly introduce the FB-LTAG formalism and XTAG grammar in this section.

3.1 Feature-Based Lexicalized Tree Adjoining Grammars

FB-LTAG is based on tree adjoining grammar (TAG) proposed in (Joshi *et al*., 1975). The TAG formalism is a formal tree rewriting system, which consists of a set of elementary trees, corresponding to minimal linguistic structures that localize the dependencies, such as specifying the predicate-argument structure of a lexeme. Elementary trees are divided into

initial and auxiliary trees. Initial trees are those for which all non-terminal nodes on the frontier are substitutable, marked with "↓". Auxiliary trees are defined as initial trees, except that exactly one frontier, nonterminal node must be a foot node, marked with "*", with the same label with the root node. Two operations - substitution and adjunction are provided in TAG to adjoin elementary trees.

FB-LTAG has two important characteristics: First, it is a lexicalized TAG (Schabes, 1988). Thus each elementary tree is associated with at least one lexical item. Second, it is a feature-based lexicalized TAG (Vijay-Shanker & Joshi, 1988). Each node in an elementary tree is constrained by two sets of feature-value pairs (two AVMs). One AVM (top AVM) defines the relation of the node to its super-tree, and the other AVM (bottom AVM) defines the relation of the node to its descendants. We use Fig1 and Fig2[1] to illustrate the substitution and adjunction operations with the unification framework respectively.

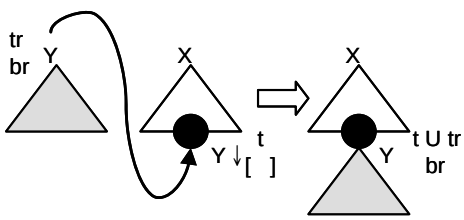


Figure 1. Substitution of FB-LTAG

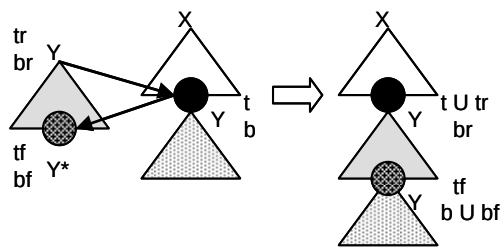


Figure 2. Adjunction of FB-LTAG

In Fig 1, we can see that the feature structure of a new node created by substitution inherits the union of the features of the original nodes. The top feature of the new node is the union of the top features of the two original nodes, while the bottom feature of the new node is simply the bottom feature of the top node of the substituting tree. In Fig 2, we can see that the node being adjoined into splits, and its top feature unifies with the top feature of the root adjoining node, while its bottom feature unifies with the bottom feature of the foot adjoining node.

3.2 XTAG English Grammar

XTAG English grammar (XTAG group, 2001) is designed using the FB-LTAG formalism, released[2] by UPENN in 2001. The range of syntactic phenomena that can be handled is large. It defines 57 major elementary trees (tree families) and 50 feature types, such as agreement, case, mode (mood), tense, passive, etc, for its 20027 lexical entries. Each lexical entry is

[1] The two figures and their descriptions are based on the XTAG technical report (XTAG group, 2001)

[2] http://www.cis.upenn.edu/~xtag/gramrelease.html

associated with at least one elementary tree, and each elementary tree is associated with at least one AVM. For example, Fig 3 shows the simplified elementary tree of "saw". "<number>" indicates the same feature value. For example, the feature – "arg_3rdsing" in bottom AVM of root S should have the same feature value of "arg_3rdsing" in top AVM of VP. In our implementation, it is coded using the same object in an object-oriented programming language. Since the feature value of mode in top AVM of "S ↓ " is "base", we know that "saw" can only be followed by a sentence with a base verb. For example, "He saw me do that" shown in Fig 4(a) is a grammatical sentence while "He saw me to do that" shown in Fig 4(b) is an ungrammatical sentence because "saw" is not allowed to be followed by an infinitive sentence.

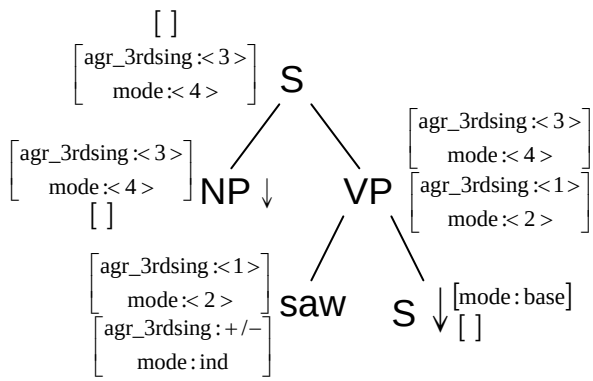


Figure 3. Elementary tree for "saw"

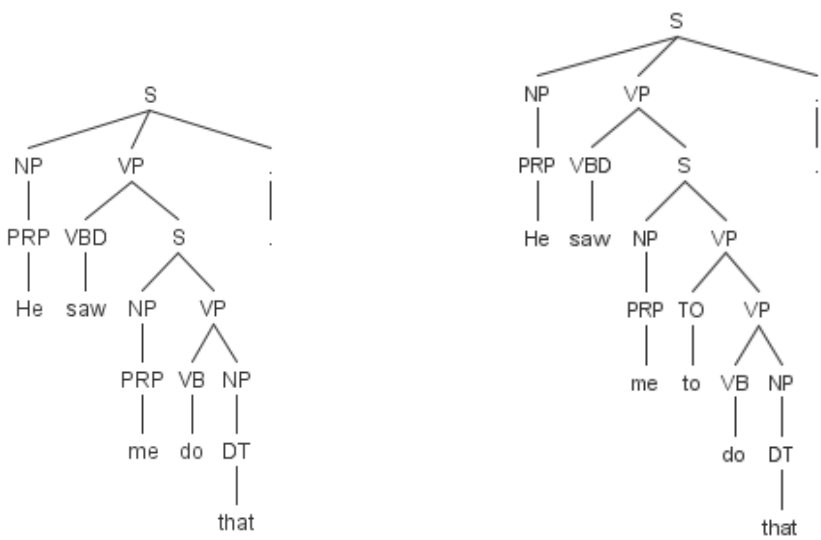


Figure 4(a). Grammatical sentence of "saw" (b) Ungrammatical sentence of "saw"

But if we look at the simplified elementary tree of "asked" shown in Fig 5, we can find that "asked" can only be followed by a sentence with an infinitive sentence (inf). For example, "He asked me to do that" shown in Fig 6(a) is a grammatical sentence while "He asked me do that" shown in Fig 6(b) is an ungrammatical sentence because "asked" is not allowed to be followed by a sentence with a base verb.

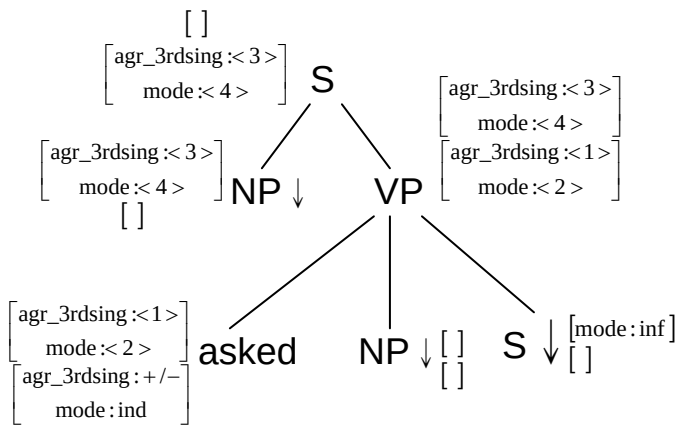


Figure 5. Elementary tree for "asked"

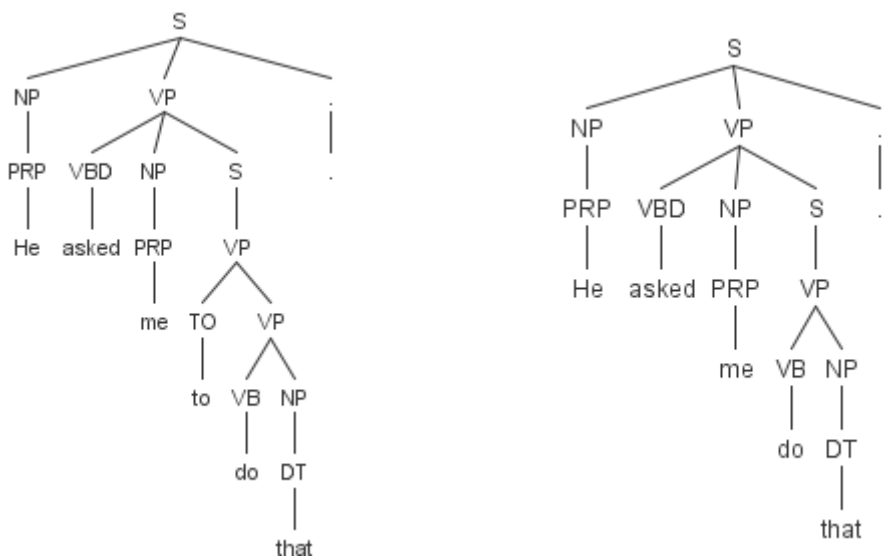


Figure 6(a). Grammatical sentence of "asked"(b) Ungrammatical sentence of "asked"

4. Syntactic Error Detection

Our procedure for syntactic error detection includes 1. decomposing each sentence hypothesis parse tree into elementary trees, 2. associating each elementary tree with AVMs through look-up in the XTAG grammar, and 3. reconstructing the original parse tree out of the elementary trees using substitution and adjunction operations along with AVM unifications.

When unification of the AVMs fails, a grammatical error has been detected and its error type is also identified by the corresponding feature in the AVM. In order to simultaneously detect multiple error types and their corresponding words, we adjust the traditional unification definition to allow the unification procedure to continue after an AVM failure occurs and also propagate the failure information to relevant words. We call the modified unification *fail propagation unification*.

Each step is illustrated in this section.

4.1 Decomposing to Elementary trees

Given a translation sentence, we first get its syntactic parse using the Stanford parser (Klein & Manning, 2003) and then decompose the parse to multiple elementary trees by using an elementary tree extractor, a modification of (Chen & Vijay-Shanker, 2000). After that, each lexical item in the sentence will be assigned one elementary tree. Taking the sentence – "Many young student play basketball" as an example, its parse and extracted elementary trees are shown in Fig 7 and Fig 8, respectively. In Fig 8, the arrows represent relations among the elementary trees and the relations are either substitution or adjunction. In this example, the two upper arrows are substitutions and the two bottom arrows are adjunctions.

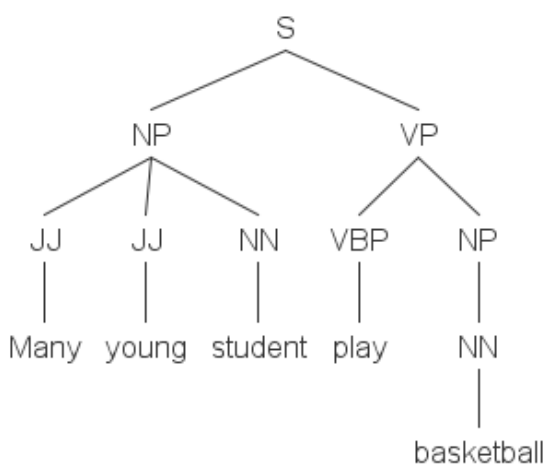


Figure 7. Parse of "Many young student play basketball"

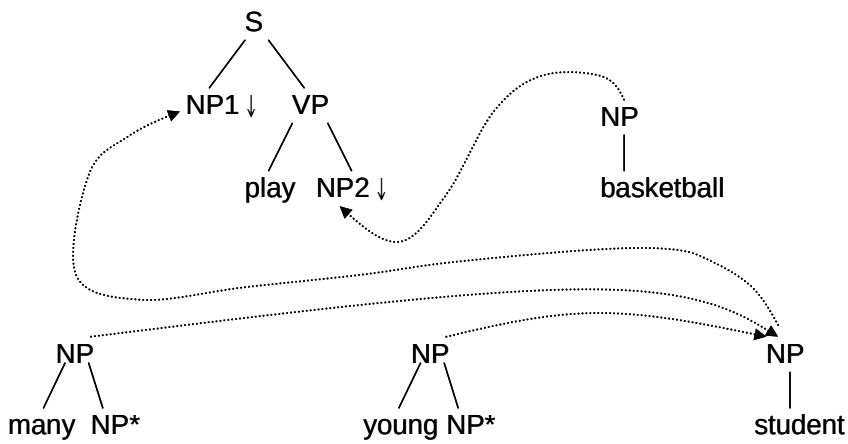


Figure 8. The elementary trees of 'Many young student play basketball" and their relations

4.2 Associating AVMs to Elementary trees

Each elementary tree is associated with AVMs through look-up in the XTAG English grammar. Using the same example of the sentence – "Many young student play basketball", its elementary trees, relations and one set of AVMs (simplified version) are shown in Fig 9. To keep tracing what word(s) that a feature value relates to for the next step of reconstruction, we design a new data structure of word set, named "word trace". It is represented by "{…}" and attached with each feature value except the value of "null", such as "agr_num:pl{play}" in Fig 9.

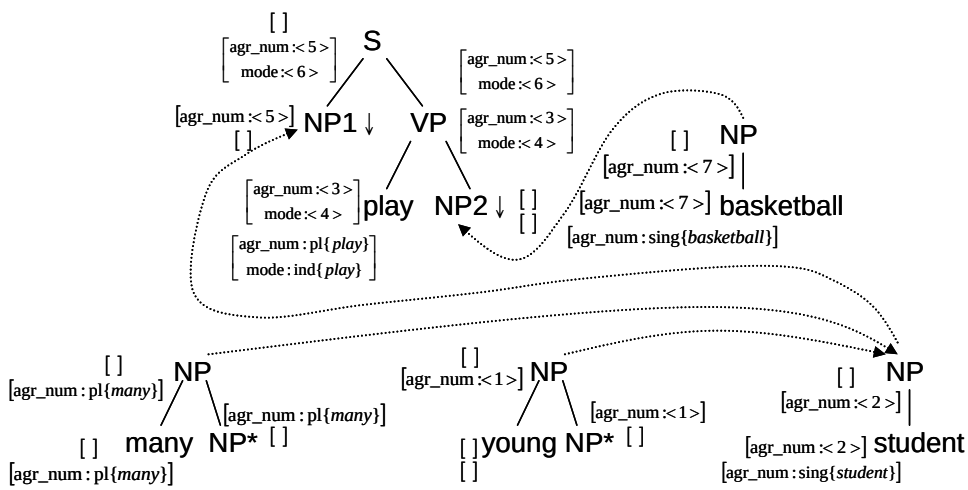


Figure 9. The elementary trees of 'Many young student play basketball", their relations and AVMs (simplified version).

When we loop up the XTAG English Grammar, sometimes one elementary tree could have multiple possible AVM associations. For example, for the verb “are”, one of its elementary trees is associated with three different AVMs, one for 2nd person singular, one for 2nd person plural, and one for 3rd person plural. Unless we can reference the context for “are” (e.g., its subject), we are not sure which AVM should be used in the reconstruction. So we postpone this decision until later in the reconstruction process. At this point, we associate each elementary tree with its all possible AVMs defined in the XTAG English Grammar.

4.3 Reconstruction Framework

Once the elementary trees are associated with AVMs, they will be used to reconstruct the original parse tree through substitution and adjunction operations which are indicated during the process of decomposing a parse tree to elementary trees. The reconstruction process is able to decide if there is any conflict with the AVMs values. When a conflict occurs, it will cause an AVM unification failure, referring to a certain grammatical error.

We already illustrated how substitution and adjunctions along with AVM unifications work in section 3.1; one implementation complement is, once the original parse is constructed, it is necessary to unify every node’s top and bottom AVMs in the constructed tree. This is because, in XTAG grammar, most AVM values are assigned in the anchor nodes of elementary trees and were not unified with others yet. This end step will assure that all related AVMs are unified.

As we stated in Section 4.2, sometimes we are not sure which AVM association for one elementary tree should be used in the reconstruction. So our strategy is to carry out reconstruction process for all sets out of every elementary tree’s each possible AVM association. We choose the set that causes the minimal grammatical errors as the detection result.

4.4 Fail Propagation Unification

Our system detects grammatical errors by identifying unification fails. However, traditional unification does not define how to proceed after fails occur, and also lacks an appropriate structure to record error traces. So we extend it as follows:

[f=x] {t_1}	U	[f=x] {t_2}	=>	[f=x] {t_1} union {t_2}	(1)
[f=x] {t_1}	U	[f=null]	=>	[f=x] {t_1}	(2)
[f=null]	U	[f=null]	=>	[f=null]	(3)
[f=x] {t_1}	U	[f=y] {t_2}	=>	[f=fail] {t_1} union {t_2}	(4)
[f=fail] {t_1}	U	[f=null]	=>	[f=fail] {t_1}	(5)
[f=fail] {t_1}	U	[f=y] {t_2}	=>	[f=fail] {t_1} union {t_2}	(6)
[f=fail] {t_1}	U	[f=fail] {t_2}	=>	[f=fail] {t_1} union {t_2}	(7)

Where f is a feature type, such as “arg_num”; x and y are two different feature values; U represents the “unify” operation; t1 and t2 are word traces introduced in section 4.2. “fail” is also defined as a kind of value.

(1)~(4) are actually traditional unification definitions except that the word trace union operations and the characteristic of fail have been added. When a unification failure occurs in (4), the unification procedure does not halt but only assigns f a value of fail and proceeds. (5)~(7) propagate the fail value to the related words’ AVMs. We use the following two unifications occurring in order in Fig 9’s adjoining operations to illustrate the procedure of fail propagation unification:

[arg_num=pl]{many} U [arg_num=sing]{student}
=> [arg_num =fail]{many,student}

[arg_num=fail]{many, student} U [arg_num=pl]{play}
=> [arg_num =fail]{many,student,play}

By the feature value of “fail” and the word trace, we identify that there is an agr_num error related to three words – “many”, “student” and “play”.

All AVMs in Fig 9 after unifications along with reconstruction operations are shown in Fig 10.

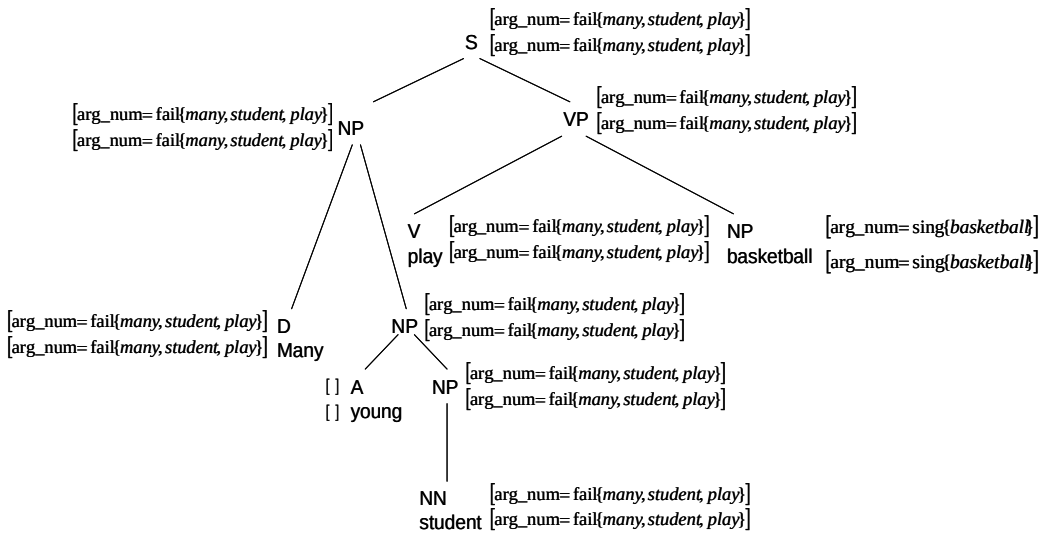


Figure 10. Reconstructed parse of the sentence- "Many young student play basketball" after unifications with fail propagation

5. Syntactic Error Correction

Once error types and their corresponding words are detected, one is able to correct errors based on an unified consideration of all related words under the same error types.

Given a set of related ungrammatical words, there are two tasks for the correction process: which words should be corrected and how to correct them? To date, we have developed the following simple mechanism to handle the agreement problem: First, the words whose feature value is in the minority will be selected to be corrected. We call this feature-value voting. Take the above example: "student" should be corrected since its agr_num is "sing" and the other two words' agr_num is "plural". When facing cases of equal votes, we tend to correct nouns if there are nouns.

Once the corrected words are selected, we replace them with their variations but with the same elementary tree type, such as replacing the above "student" with "students."

6. Experiment

Among the 57 major elementary trees and 50 feature types that XTAG defines, we have implemented 26 major elementary trees and 4 feature types – agr_pers, arg_num, arg_3rdsing and several cases of mode/mood at this point (The first three belong to agreement features.) We apply our syntactic error detection and correction on 422 translation sentences of six Chinese-English machine translation systems A~F from the DARPA Global Autonomous Language Exploitation (GALE) 2008 evaluation. Every source sentence is provided along

with four target references. The six systems are described in Table 1, and the results of syntactic error detection for agreement and mode errors and correction for agreement errors are shown in Table 2.

Table 1. Six MT systems

	System name	Approach
A	NRC	phrase-based SMT
B	RWTH-PBT	phrase-based SMT
C	RWTH-PBT-AML	phrase-based SMT + source reordering
D	RWTH-PBT-JX	phrase-based SMT + Chinese word segmentation
E	RWTH-PBT-SH	phrase-based SMT + source reordering + rescoring
F	SRI-HPBT	hierarchical phrase-based SMT

Table 2. The results of syntactic error detection and correction

	Detected sentences (arg error + mode error)	Corrected sentences (arg error)	Bleu for all sentences (before)	Bleu for all sentences (after)	Bleu for corrected sentences (before)	Bleu for corrected sentences (after)
A	23	9	32.99	32.99	26.75	27.80
B	23	14	27.95	27.97	22.08	23.03
C	18	7	34.40	34.41	32.13	32.67
D	25	14	32.96	32.99	31.49	32.17
E	30	11	34.64	34.68	29.31	30.61
F	18	8	34.13	34.14	29.15	28.83

From Table 2, even the overall Bleu score for all sentences is not significantly improved, but if we take a close look at those corrected sentences for agreement errors and calculate their Bleu scores, we can see the corrected translations are improved for every system except for one (F), which shows the effectiveness and potential of our approach.

7. Conclusion

This paper presents a new FB-LTAG-based syntactic error detection and correction mechanism along with a novel AVM unification method to simultaneously detect multiple ungrammatical types and their corresponding words for machine translation. The mechanism can also be applied to other languages if the grammar is well defined in the FB-LTAG structure of certain languages.

While the basic design philosophy and algorithm are fully described in this paper, we are continuing to implement more elementary trees and feature types defined in the XTAG grammar, and we are extending our correction mechanism as our future work.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. This work is supported by the National Science Foundation via Grant No. 0910778 entitled "Richer Representations for Machine Translation". All views expressed in this paper are those of the authors and do not necessarily represent the view of the National Science Foundation.

Reference

Alam, M. J., UzZaman, N. & Khan, M. (2006). N-gram based Statistical Grammar Checker for Bangla and English. In *Proceedings of ninth International Conference on Computer and Information Technology (ICCIT 2006)*, Dhaka, Bangladesh.

Atwell, E. S. & Elliot, S. (1987). Dealing with Ill-formed English Text. In: R. Garside, G. Leech and G. Sampson (Eds.) *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Chen, J. & Vijay-Shanker, K. (2000). Automated extraction of TAGs from the Penn treebank. In *Proceedings of the Sixth International Workshop on Parsing Technologies*.

Heidorn, G. E. (2000). Intelligent writing assistance. In R. Dale, H. Moisl and H. Somers (eds.), *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*. Marcel Dekker, New York. 181-207.

Huang, A., Kuo, T. T., Lai, Y. C. & Lin, S. D. (2010). Identifying Correction Rules for Auto Editing. In *Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING)*, 251-265.

Jensen, K., Heidorn, G. E. & Richardson, S. D. (Eds.) (1993). *Natural Language Processing: The PLNLP Approach*, Kluwer Academic Publishers.

Joshi, A. K., Levy, L. S. & Takahashi M. (1975). Tree Adjunct Grammars. *Journal of Computer and System Science*, 10, 136-163.

Klein, D. & Manning, C. D. (2003). Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423-430.

Ma, W. Y. & McKeown, K. (2011). System Combination for Machine Translation Based on Text-to-Text Generation. In *Proceedings of Machine Translation Summit XIII*. Xiamen, China.

Naber, D. (2003). A Rule-Based Style and Grammar Checker. *Diploma Thesis*. University of Bielefeld, Germany.

Schabes, Y., Abeille, A. & Joshi, A. K. (1988). Parsing strategies with 'lexicalized' grammars: Application to tree adjoining grammars. In *Proceeding of 12th International Conference on Computational Linguistics (COLING'88)*, Budapest, Hungary.

Stymne, S. & Ahrenberg, L. (2010). Using a Grammar Checker for Evaluation and Postprocessing of Statistical Machine Translation. In *Proceedings of International Conference on Language Resources and Evaluation (LREC).*

Vijay-Shanker, K. & Joshi, A. K. (1988). Feature structure based tree adjoining grammar, In *Proceeding of 12th International Conference on Computational Linguistics (COLING'88)*, 714-719.

Wu, S. H., Su, C. Y., Jiang, T. J. & Hsu, W. L. (2006). An Evaluation of Adopting Language Model as the Checker of Preposition Usage. In *Proceedings of the Conference on Computational Linguistics and Speech Processing (ROCLING).*

XTAG Group. (2001). A Lexicalized Tree Adjoining Grammar for English. *Technical Report IRCS 01-03*, University of Pennsylvania.

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 4, December 2012, pp.15-32

TQDL: Integrated Models for Cross-Language Document Retrieval

Long-Yue WANG*, Derek F. WONG*, and Lidia S. CHAO*

Abstract

This paper proposed an integrated approach for Cross-Language Information Retrieval (CLIR), which integrated with four statistical models: **T**ranslation model, **Q**uery generation model, **D**ocument retrieval model and **L**ength Filter model. Given a certain document in the source language, it will be translated into the target language of the statistical machine translation model. The query generation model then selects the most relevant words in the translated version of the document as a query. Instead of retrieving all the target documents with the query, the length-based model can help to filter out a large amount of irrelevant candidates according to their length information. Finally, the left documents in the target language are scored by the document searching model, which mainly computes the similarities between query and document.

Different from the traditional parallel corpora-based model which relies on IBM algorithm, we divided our CLIR model into four independent parts but all work together to deal with the term disambiguation, query generation and document retrieval. Besides, the TQDL method can efficiently solve the problem of translation ambiguity and query expansion for disambiguation, which are the big issues in Cross-Language Information Retrieval. Another contribution is the length filter, which are trained from a parallel corpus according to the ratio of length between two languages. This can not only improve the recall value due to filtering out lots of useless documents dynamically, but also increase the efficiency in a smaller search space. Therefore, the precision can be improved but not at the cost of recall.

* Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory, Department of Computer and Information Science, University of Macau, Macau S. A. R., China
E-mail: vincentwang0229@hotmail.com
The author for correspondence is Long-Yue Wang.

In order to evaluate the retrieval performance of the proposed model on cross-languages document retrieval, a number of experiments have been conducted on different settings. Firstly, the Europarl corpus which is the collection of parallel texts in 11 languages from the proceedings of the European Parliament was used for evaluation. And we tested the models extensively to the case that: the lengths of texts are uneven and some of them may have similar contents under the same topic, because it is hard to be distinguished and make full use of the resources.

After comparing different strategies, the experimental results show a significant performance of the method. The precision is normally above 90% by using a larger query size. The length-based filter plays a very important role in improving the F-measure and optimizing efficiency.

This fully illustrates the discrimination power of the proposed method. It is of a great significance to both cross-language searching on the Internet and the parallel corpus producing for statistical machine translation systems. In the future work, the TQDL system will be evaluated for Chinese language, which is a big changing and more meaningful to CLIR.

Keywords: Cross-Language Document Retrieval, Statistical Machine Translation, TF-IDF, Document Translation-Based, Length-Based Filter.

1. Introduction

With the flourishing development of the Internet, the amount of information from a variety of domains is rising dramatically. Especially after the advent of the World Wide Web (WWW) in the 1900s, the amount of online information from the government, scientific and business communities has risen dramatically. Although much word has been done to develop effective and efficient retrieval systems for monolingual resources, the diversity and the explosive growth of information in different languages drove a great need for information retrieval that could cross language boundaries (Ballesteros *et al.*, 1988).

The issues of CLIR have been discussed for several decades. Its task addresses a situation in which a user tries to search a set of documents written in one language using a query in a different language (Kishida, 2005). It is of great significance, allowing people access information resources written in non-native languages and aligning documents for statistical machine translation (SMT) systems, of which quality is heavily dependent upon the amount of parallel sentences used in constructing the system.

In this paper, we focus on the problems of translation ambiguity, query generation and searching score which are keys to the retrieval performance. First of all, in order to increase the probability that the best translation can be selected from multiple ones, which occurs in the

target documents, the context and the most likely probability of the whole sentence should be considered. So we apply document translation approach using SMT model instead of query translation, although the latter one may require fewer computational resources. After the source documents are translated into the target language, the problem is transformed from bilingual environment to monolingual one, where conventional IR techniques can be used for document retrieval. Secondly, some terms in a certain document will be selected as query, which can distinguish the document from others. However, some of the words occur too frequently to be useful, which cannot distinguish target documents. This mostly includes two cases: one is that the word frequency is high in all the documents of a set, which is usually classified as stop word; the other one is that the frequency is moderate in several documents of a set. These words are poor in the ability of distinguishing documents. Thus, the query generation model should pick the words that occur more frequently in a certain document while less frequently in other documents. Finally, the document searching model evaluates the similarity between the query and each document. This model should give a higher score to the target document which covers the most relevant words in the given query. However, another problem is that word overlap between a query and a wrong document is more probable when the document and the query are expressed in the same language. For example, Document *A* is larger and contains another smaller document *B*. So the retrieval system would be confused with a query including the information of *B*. In order to solve this problem, the length ratio of a language pair is considered. As the search space is reduced, both the speed efficiency and the recall value will be improved clearly.

There are two cases to be considered when we investigated the method. In one case, the lengths of documents are uneven, which are hard to balance the scores between large and small documents. In the other case, the contents of the documents are very similar, which are not easy to distinguish for retrieval. The results of experiments reveal that the proposed model shows a very good performance in dealing with both cases.

The paper is organized as follows. The related works are reviewed and discussed in Section 2. The proposed CLIR approach based on statistical models is described in Section 3. The resources and configurations of experiments for evaluating the system are detailed in Section 4. Results, discussion and comparison between different strategies are given in Section 5 followed by a conclusion and future improvements to end the paper.

2. Related Work

The issues of CLIR have been discussed from different perspectives for several decades. In this section, we briefly describe some related methods.

From a statistical perspective, the CLIR problem can be treated as document alignment. Given a set of parallel documents, the alignment that maximizes the probability over all

possible alignments is retrieved (Gale & Church, 1991) as follows:

$$\arg\max \Pr(A \mid D_s, D_t) \approx \arg\max_{A} \prod_{(L_s \Leftrightarrow L_t) \in A} \Pr(L_s \leftrightarrow L_t \mid L_s L_t) \tag{1}$$

where A is an alignment, D_s and D_t are the source and target documents, respectively L_1 and L_2 are the documents of two languages, $L_s \leftrightarrow L_t$ is an individual aligned pairs, an alignment A is a set consisting of $L_s \leftrightarrow L_t$ pairs.

On the matching strategies for CLIR, query translation is most widely used method due to its tractability (Gao *et al*., 2001). However, it is relatively difficult to resolve the problem of term ambiguity because "queries are often short and short queries provide little context for disambiguation" (Oard & Diekema, 1998). Hence, some researchers have used document translation method as the opposite strategies to improve translation quality, since more varied context within each document is available for translation (Braschler & Schauble, 2001; Franz *et al*., 1999).

However, another problem introduced based on this approach is word (term) disambiguation, because a word may have multiple possible translations (Oard & Diekema, 1998). Significant efforts have been devoted to this problem. Davis and Ogden (1997) applied a part-of-speech (POS) method which requires POS tagging software for both languages. Marcello et al. presented a novel statistical method to score and rank the target documents by integrating probabilities computed by query-translation model and query-document model (Federico & Bertoldi, 2002). However, this approach cannot aim at describing how users actually create queries which have a key effect on the retrieval performance. Due to the availability of parallel corpora in multiple languages, some authors have tried to extract beneficial information for CLIR by using SMT techniques. Sánchez-Martínez *et al*. (Sánchez-Martínez & Carrasco, 2011) applied SMT technology to generate and translate queries in order to retrieve long documents.

Some researchers like Marcello, Sánchez-Martínez *et al*. have attempted to estimate translation probability from a parallel corpus according to a well-known algorithm developed by IBM (Brown *et al*., 1993). The algorithm can automatically generate a bilingual term list with a set of probabilities that a term is translated into equivalents in another language from a set of sentence alignments included in a parallel corpus. The IBM Model 1 is the simplest among the five models and often used for CLIR. The fundamental idea of the Model 1 is to estimate each translation probability so that the probability represented is maximized

$$P(t \mid s) = \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} P(t_j \mid s_i) \tag{2}$$

where t is a sequence of terms $t_1, \ldots, t_m$ in the target language, s is a sequence of terms $s_1, \ldots, s_l$ in the source language, $P(t_j|s_i)$ is the translation probability, and ε is a parameter ($\varepsilon = P(m|e)$),

where *e* is target language and *m* is the length of source language). Eq. (2) tries to balance the probability of translation, and the query selection, in which problem still exists: it tends to select the terms consisting of more words as query because of its less frequency, while cutting the length of terms may affect the quality of translation. Besides, the IBM model 1 only proposes translations word-by-word and ignores the context words in the query. This observation suggests that a disambiguation process can be added to select the correct translation words (Oard & Diekema, 1998). However, in our method, the conflict can be resolved through contexts.

If translated sentences share cognates, then the character lengths of those cognates are correlated (Yang & Li, 2004). Brown *et al*. (1991) and Gale and Church (1991) have developed the models based on relationship between the lengths of sentences that are mutual translations. Although it has been suggested that length-based methods are language-independent (Gale & Church, 1991), they really rely on length correlations arising from the historical relationships of the languages being aligned.

The length-based model assumes that each term in L_s is responsible for generating some number of terms in L_t. This leads to a further approximation that encapsulates the dependence to a single parameter δ. $\delta(l_s, l_t)$ is function of l_s and l_t, which can be designed according to different language pairs. The length-based method is developed based on the following approximation to Eq. (3):

$$\Pr(L_s \leftrightarrow L_t \mid L_s, L_t) \approx \Pr(L_s \leftrightarrow L_t \mid \delta(l_s, l_t)) \tag{3}$$

3. Proposed Models

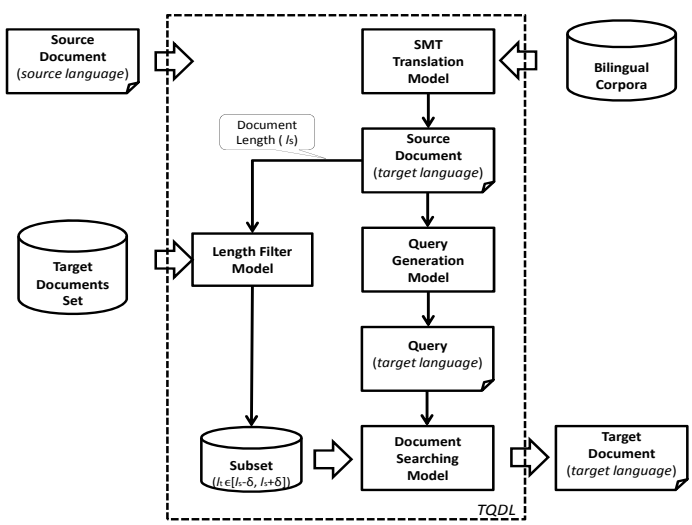


Figure 1. The proposed approach for CLIR

The approach relies on four models: translation model which generates the most probable translation of source documents; query generation model which determines what words in a document might be more favorable to use in a query; length filter model dynamically create a subset of candidates for retrieval according to the length information; and document searching model, which evaluates the similarity between a given query and each document in the target document set. The workflow of the approach for CLIR is shown in Fig. 1.

3.1 Translation Model

Currently, the good performing statistical machine translation systems are based on phrase-based models which translate small word sequences at a time. Generally speaking, translation model is common for contiguous sequences of words to translate as a whole. Phrasal translation is certainly significant for CLIR (Ballesteros & Croft, 1997), as stated in Section 1. It can do a good job in dealing with term disambiguation.

In this work, documents are translated using the translation model provided by Moses, where the log-linear model is considered for training the phrase-based system models (Och & Ney, 2002), and is represented as:

$$p(e_1^I \mid f_1^J) = \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J))}{\sum_{e'^I_1} \exp(\sum_{m=1}^{M} \lambda_m h_m(e'^I_1, f_1^J))} \tag{4}$$

where h_m indicates a set of different models, λ_m means the scaling factors, and the denominator can be ignored during the maximization process. The most important models in Eq. (4) normally are phrase-based models which are carried out in source to target and target to source directions. The source document will maximize the equation to generate the translation including the words most likely to occur in the target document set.

3.2 Query Generation Model

After translating the source document into the target language of the translation model, the system should select a certain amount of words as a query for searching instead of using the whole translated text. It is for two reasons, one is computational cost, and the other is that the unimportant words will degrade the similarity score. This is also the reason why it often responses nothing from the search engines on the Internet when we choose a whole text as a query.

In this paper, we apply a classical algorithm which is commonly used by the search engines as a central tool in scoring and ranking relevance of a document given a user query. Term Frequency–Inverse Document Frequency (TF-IDF) calculates the values for each word in a document through an inverse proportion of the frequency of the word in a particular

document to the percentage of documents where the word appears (Ramos, 2003). Given a document collection D, a word w, and an individual document $d \in D$, we calculate

$$P(w,d) = f(w,d) \times \log \frac{|D|}{f(w,D)} \tag{5}$$

where $f(w, d)$ denotes the number of times w that appears in d, $|D|$ is the size of the corpus, and $f(w,D)$ indicates the number of documents in which w appears in D (Berger *et al*., 2000).

In implementation, if w is an Out-of-Vocabulary term (OOV), the denominator $f(w,D)$ becomes zero, and will be problematic (divided by zero). Thus, our model makes *log* ($|D|/f(w,D)$)=1 (IDF=1) when this situation occurs. Additionally, a list of stop-words in the target language is also used in query generation to remove the words which are high frequency but less discrimination power. Numbers are also treated as useful terms in our model, which also play an important role in distinguishing the documents. Finally, after evaluating and ranking all the words in a document by their scores, we take a portion of the (n-best) words for constructing the query and are guided by:

$$Size_q = [\lambda_{percent} \times Len_d] \tag{6}$$

$Size_q$ is the number of terms. $\lambda_{percent}$ is the percentage and is manually defined, which determines the $Size_q$ according to Len_d, the length of the document. The model uses the first $Size_q$-th words as the query. In another word, the larger document, the more words are selected as the query.

3.3 Document Retrieval Model

In order to use the generated query for retrieving documents, the core algorithm of the document retrieval model is derived from the Vector Space Model (VSM). Our system takes this model to calculate the similarity of each indexed document according to the input query. The final scoring formula is given by:

$$Score(q,d) = coord(q,d) \sum_{t_{in}q} tf(t,d) \times idf(t) \times bst \times norm(t,d) \tag{7}$$

where $tf(t,d)$ is the term frequency factor for term t in document d, $idf(t)$ is the inverse document frequency of term t, while $coord(q,d)$ is frequency of all the terms in query occur in a document. bst is a weight for each term in the query. $Norm(t,d)$ encapsulates a few (indexing time) boost and length factors, for instance, weights for each document and field. As a summary, many factors that could affect the overall score are taken into account in this model.

3.4 Length Filter Model

In order to obtain a suitable filter, we firstly analyzed the golden data[1] of ACL Workshop on SMT 2011, which includes Spanish, English, French, German and Czech 5 languages and 10 language pairs. English-Spanish language pair was used for analyzing and the data of the corpus are summarizes in Table 1.

Table 1. Analytical Data of Corpus of ACL Workshop on SMT 2011

Dataset	Size of corpus		
	No. of Sentences	***No. of Characters***	***Ave. No. Characters***
English	3,003	74,753	25
Spanish	3,003	79,426	26

Fig. 2 plots the distribution of word number in each aligned sentences. l_t is the length of English sentence while l_s is the length of sentence in Spanish. So the expectation is c= E (l_t/l_s) =1.0073, with the correlation R^2 = 0.9157. This shows that the data points are not substantially scatter in the plot and many data points are along with the regression line. Therefore, it is suitable to design a filter based on length ratio.

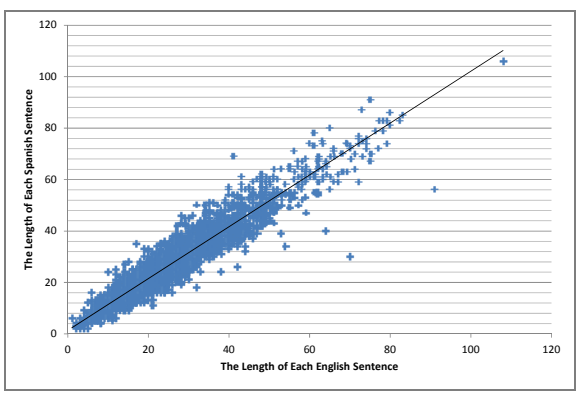


Figure 2. The length ratio of Spanish-English sentences.

To obtain an estimated length-threshold (δ) for filter model, the function δ (l_s, l_t) can be designed as follows:

$$\delta(l_s, l_t) = \frac{|l_t - l_s|}{l_s} \tag{8}$$

where l_s and l_t respectively stand for the length of a certain aligned sentence in the corpus we used. Finally, we got the average δ of around **0.15**. In implementation, we choose 4δ instead of δ to avoid some unnormal cases, where the right document would be discarded by the filter.

[1] It can be download from http://www.statmt.org/wmt11/

Filter F describes the relation between bilingual sentences based on the length ratio. Since western languages are similar in terms of word representation, the length ratio can be simply estimated as a 1:1. Given a certain document in source language, F can collect a subset for retrieval according to the average length ratio. So F is designed as follows:

$$F = \begin{cases} 1, length_t \in C \\ 0, length_t \notin C \end{cases}, C = [length_s - \delta, length_s + \delta] \tag{9}$$

where $length_s$ is the length of source document, and $length_t$ is the length of target document. δ is an average threshold obtained through Eq. (8), C is a confidence interval. If $length_t$ is included in C, F is 1, which has a chance to be retrieved, otherwise set as 0, which will be skipped during searching.

4. Model Evaluation

4.1 Datasets

In order to evaluate the retrieval performance of the proposed model on text of cross languages, we use the Europarl corpus[2] which is the collection of parallel texts in 11 languages from the proceedings of the European Parliament (Koehn, 2005). The corpus is commonly used for the construction and evaluation of statistical machine translation. The corpus consists of spoken records held at the European Parliament and are labeled with corresponding IDs (e.g. <CHAPTER *id*>, <SPEAKER *id*>). The corpus is quite suitable for use in training the proposed probabilistic models between different language pairs (e.g. English-Spanish, English-French, English-German, etc.), as well as for evaluating retrieval performance of the system.

Table 2. Analytical Data of Corpus

Dataset	Size of corpus			
	Documents	**Sentences**	**Words**	**Ave. words in document**
Training Set	2,900	1,902,050	23,411,545	50
TestSet	23,342	80,000	7,217,827	309

The datasets (training and test set) are collected for this evaluation. The chapters from April 1998 to October 2006 were used as a training set for model construction, both for training the Language Model (LM) and Translation Model (TM). While the chapters from April 1996 to March 1998 were considered as the testing set for evaluating the performance of the model. Besides, each paragraph (split by <SPEAKER *id*> label) is treated as a document, for dealing with the low discrimination power. The analytical data of the corpus are presented

[2] Available online at http://www.statmt.org/europarl/.

in Table 2. The TestSet contains 23,342 documents, of which length is 309 in average. Actually 30% of documents are much more or less than the average number. Table 1 summarizes the number of documents, sentences, words and the average word number of each document.

4.2 Evaluation Metrics

The most frequent and basic evaluation metrics for information retrieval are precision and recall, which are defined as follows (Manning *et al*., 2008):

$$\mathrm{Pr}ecison = \frac{\#(relevant\quad items\quad retrieved)}{\#(retrieved\quad items)} \tag{10}$$

$$\mathrm{Re}call = \frac{\#(relevant\quad items\quad retrieved)}{\#(relevant\quad items)} \tag{11}$$

For reporting the evaluation of our method, we used the *F*1 measure, the recall and the precision values. *F*1-measure (*F*) is formulated by Van Rijsbergen as a combination of recall (*R*) and precision (*P*) with an equal weight in the following form:

$$F = \frac{2PR}{P+R} \tag{12}$$

4.3 Experimental Setup

In order to evaluate our proposed model, the following tools have been used.

The probabilistic LMs are constructed on monolingual corpora by using the SRILM (Stolcke *et al*., 2002). We use GIZA++ (Och & Ney, 2003) to train the word alignment models for different pairs of languages of the Europarl corpus, and the phrase pairs that are consistent with the word alignment are extracted. For constructing the phrase-based statistical machine translation model, we use the open source Moses (Koehn *et al*., 2007) toolkit, and the translation model is trained based on the log-linear model, as given in Eq. (4). The workflow of constructing the translation model is illustrated in Fig. 3 and it consists of the following main steps[3]:

(1) Preparation of aligned parallel corpus.

(2) Preprocessing of training data: tokenization, case conversion, and sentences filtering where sentences with length greater than fifty words are removed from the corpus in order to comply with the requirement of Moses.

(3) A 5-gram LM is trained on Spanish data with the SRILM toolkits.

[3] See http://www.statmt.org/wmt09/baseline.html for a detailed description of MOSES training options.

(4) The phrased-based STM model is therefore trained on the prepared parallel corpus (English-Spanish) based on log-linear model of by using the nine-steps suggested in Moses.

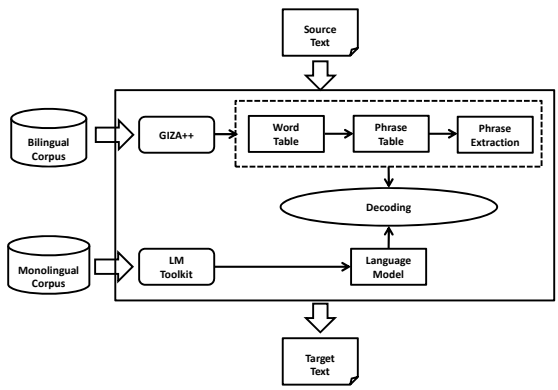


Figure 3. Main workflow of training phase

Once LM and TM have been obtained, we evaluate the proposed method with the following steps:

(1) The source documents are first translated into target language using the constructed translation model.

(2) The words candidates are computed and ranked based on a TF-IDF algorithm and the n-best words candidates then are selected to form the query based on Eq. (5) and (6).

(3) All the target documents are stored and indexed using Apache Lucene[4] as our default search engine.

(4) In retrieval, target documents are scored and ranked by using the document retrieval model to return the list of most related documents with Eq. (7).

5. Results and Discussion

A number of experiments have been performed to investigate our proposed method on different settings. In order to evaluate the performance of the three independent models, we firstly conducted experiments to test them respectively before whole the TQDL platform. The performance of the method is evaluated in terms of the ***average precision***, that is, how often the target document is included within the first N-best candidate documents when retrieved.

[4] Available at http://lucene.apache.org.

5.1 Monolingual Environment Information Retrieval

In this experiment, we want to evaluate the performance of the proposed system to retrieve documents (monolingual environment) given the query. It supposes that the translations of source documents are available, and the step to obtain the translation for the input document can therefore be neglected. Under such assumptions, the CLIR problem can be treated as normal IR in monolingual environment. In conducting the experiment, we used all of the source documents of TestSet. The steps are similar to that of the testing phase as described in Section 4.2, excluding the translation step. The empirical results based on different configurations are presented in Table 3, where the first column gives the number of documents returned against the number of words/terms used as the query.

Table 3. The average precision in Monolingual Environment

Retrieved Documents (*N*-Best)	**Query Size ($Size_q$ in %)**						
	2	**4**	**8**	**10**	**14**	**18**	**20**
1	0.794	0.910	0.993	0.989	0.986	1.000	0.989
5	0.921	0.964	1.000	1.000	1.000	1.000	0.996
10	0.942	0.971	1.000	1.000	1.000	1.000	0.996
20	0.946	0.978	1.000	1.000	1.000	1.000	0.996

The results show that the proposed method gives very high retrieval accuracy, with precision of 100%, when the top 18% of the words are used as the query. In case of taking the top 5 candidates of documents, the approach can always achieve a 100% of retrieval accuracy with query sizes between 8% and 18%. This fully illustrates the effectiveness of the retrieval model.

5.2 Translation Quality

The overall retrieval performance of the system will be affected by the quality of translation. In order to have an idea the performance of the translation model we built, we employ the commonly used evaluation metric, BLEU, for such measure. The BLEU (Bilingual Evaluation Understudy) is a classical automatic evaluation method for the translation quality of an MT system (Papineni *et al*., 2002). In this evaluation, the translation model is created using the parallel corpus, as described in Section 4. We use another 5,000 sentences from the TestSet1 for evaluation[5].

[5] See http://www.statmt.org/wmt09/baseline.html for a detailed description of MOSES evaluation options.

The BLEU value, we obtained, is **32.08**. The result is higher than that of the results reported by Koehn in his work (Koehn, 2005), of which the BLEU score is **30.1** for the same language pair we used in Europarl corpora. Although we did not use exactly the same data for constructing the translation model, the value of **30.1** was presented as a baseline of the English-Spanish translation quality in Europarl corpora.

The BLEU score shows that our translation model performs very well, due to the large number of the training data we used and the pre-processing tasks we designed for cleaning the data. On the other hand, it reveals that the translation quality of our model is good.

5.3 TQDL without Filter for CLIR

In this section, the proposed model without length filter model is tested. Table 4 presents the F-measure given by TQDL system without length filter model. As illustrated, the it can only achieve up to 94.7%, counting that the desired document is returned as the most relevant document among the candidates. Although it has achieved a very good performance in the experiments, the 6.6% of documents have been discarded in the pre-processing.

Table 4. The F-measure of our system without length filter model

Retrieved Documents (*N*-Best)	**Query Size ($Size_q$ in %)**				
	2.0	**4.0**	**6.0**	**8.0**	**10.0**
1	0.905	0.943	0.942	***0.947***	0.941
2	0.922	0.949	0.949	0.953	0.950
5	0.932	0.950	0.953	0.963	0.960
10	0.936	0.954	0.960	0.968	0.971
20	0.941	0.958	0.974	0.979	0.981

To investigate the changes of the performance with removing abnormal documents (too lager or too small), query size $Size_q$ was set as a constant value (8.0%)**,** which can achieve the best precision as shown in Table 4. We believed that the abnormal document is the main obstacle to develop the performance of the system. Therefore, we removed the documents, of which length are out of a certain threshold.

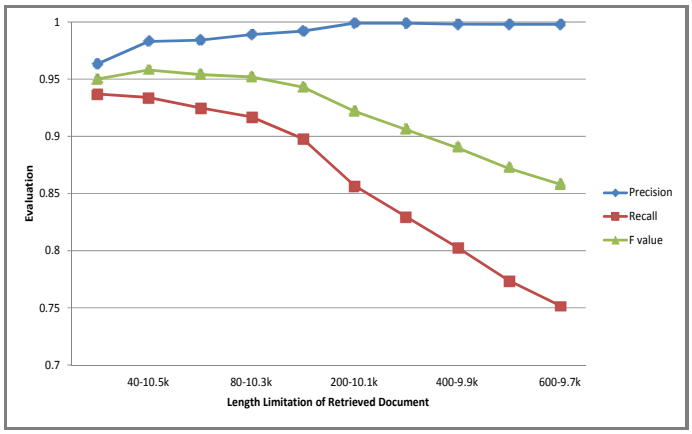


Figure 4. The changes of evaluation when removing data

Fig. 4 plots the variations of *P*, *R* and *F* with the length scope increasing. As we expected, the precision increase when the more abnormal documents are discarded from the dataset. However, the recall declines sharply, which also lead to the falling of *F*-measure. When the precision is closed to **100%**, nearly **15%** documents are removed from the dataset. So the high precision is often at the cost of reducing the recall rate. *F*-measure is only 95% at its top, so it is hard to improve the performance of CLIR using traditional methods.

5.4 TQDL with Filter for CLIR

In order to obtain a higher retrieval rate, our model has been improved from different points. Firstly, we generate the query with dynamic size, which can do better in dealing with the problem of similar documents both in length and content. In another words, the longer the document, the more words will be used for retrieval of the target documents. So the $Size_q$ is considered as a hidden variable in our document retrieval model. Besides, all the indexed documents can be filtered with *F* formula in Eq. (9), and it can alleviate the scarcity of tending to select longer documents when occurring the word overlap between shorter and longer documents, because a certain source document are only searched in a subset defined by its length. It can improve the precision without discard any so-called “abnormal” documents from dataset, so the *P*, *R* and *F* values will always be the same. Table 5 presents the *F* values given by TQDL with length filter model.

Table 5. The F-measure of our system with length filter model

Retrieved Documents (*N*-Best)	Query Size ($Size_q$ in %)				
	2.0	**4.0**	**6.0**	**8.0**	**10.0**
1	0.958	0.975	0.983	0.990	0.992
2	0.967	0.979	0.986	0.993	0.996
5	0.971	0.982	0.987	0.993	0.996
10	0.974	0.983	0.988	0.995	0.996
20	0.974	0.983	0.990	0.995	0.996

Compared with the results presented in Tables 4 and 5, it shows that the length filter model is able to give a high improvement by 4.5% in F-measure and achieve more than 99% of successful rate, in the case that the desired candidate is ranked in the first place. Above all, there is no documents waste in the dataset.

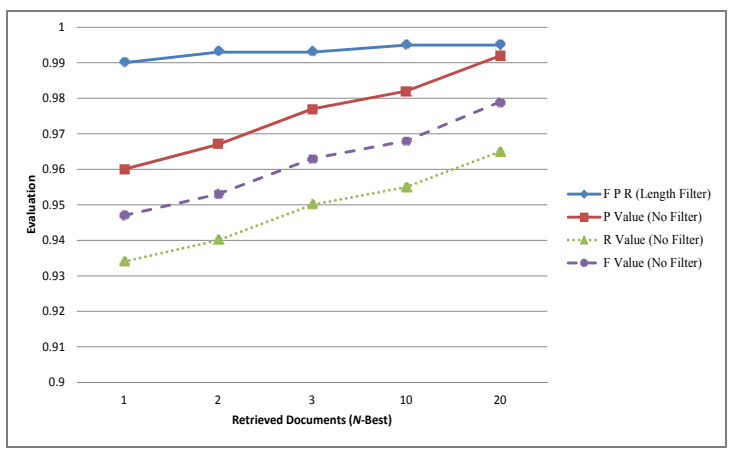


Figure 5. The changes of evaluation with N-Best

Fig. 5 presents an ideal distribution of evaluation, of which *P* and *R* should be closed to the *F* line. In this comparison, query size $Size_q$ was still set as a constant value (8.0%). With the increasing of *N*, evaluations without filter are in a low level, while the one with this filter can achieve a good and stable performance. Finally, the precision and recall values are closed to F measure, which can all keep in a high level (99%-100%).

6. Conclusion

This article presents a TQDL statistical approach for CLIR which has been explored for both large and similar documents retrieval. Different from the traditional parallel corpora-based model which relies on IBM algorithm, we divided our CLIR model into four independent parts but all work together to deal with the term disambiguation, query generation and document

retrieval. The performances showed that this method can do a good job of CLIR for not only large documents but also the similar documents. This fully illustrates the discrimination power of the proposed method. It is of a great significance to both cross-language searching on the Internet and the parallel corpus producing for statistical machine translation systems. In the future work, the TQDL system will be evaluated for Chinese language, which is a big changing and more meaningful to CLIR. In the further work, we plan to make better use of the proposed models between significantly different languages such as Portuguese-Chinese.

Acknowledgement

This work was partially supported by the Research Committee of University of Macau under grant UL019B/09-Y3/EEE/LYP01/FST, and also supported by Science and Technology Development Fund of Macau under grant 057/2009/A2.

References

Ballesteros, L., & Croft, W. B. (1988). Statistical methods for cross-language information retrieval. *Cross-language information retrieval*, 23-40.

Ballesteros, L. & Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. *ACM SIGIR Forum*, 31(SI), 84-91.

Braschler, M., & Schauble, P. (2001). Experiments with the eurospider retrieval system for clef 2000. *Cross-Language Information Retrieval and Evaluation*, 140-148.

Brown, P. F., Lai, J. C. & Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, 169-176.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D. & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2), 263-311. MIT Press.

Berger, A., Caruana, R., Cohn, D., Freitag, D. & Mittal, V. (2000). Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 192-199.

Davis, M. W. & Ogden, W. C. (1997). Quilt: Implementing a large-scale cross-language text retrieval system. *ACM SIGIR Forum*, 31(SI), 92-98.

Federico, M. & Bertoldi, N. (2002). Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 167-174.

Franz, M., McCarley, J. S. & Ward, R. T. (1999). Ad hoc, cross-language and spoken document information retrieval at IBM. NIST Special Publication: *The 8th Text Retrieval Conference* (*TREC-8*).

Gale, W. A. & Church, K. W. (1991). Identifying word correspondences in parallel texts. In *Proceedings of the workshop on Speech and Natural Language*, 152-157.

Gao, J., Nie, J. Y., Xun, E., Zhang, J., Zhou, M., & Huang, C. (2001). Improving query translation for cross-language information retrieval using statistical models.In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 96-104.

Kishida, K. (2005). Technical issues of cross-language information retrieval: a review. *Information processing & management*, 41(3), 433-455.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT summit*, 5.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., *et al*. (2007). Moses: Open source toolkit for statistical machine translation. *Annual meeting-association for computational linguistics*, 45(2), 2.

Manning, C. D., Raghavan, P. & Schütze, H. (2008). Introduction to information retrieval (Vol. 1). *Cambridge University Press Cambridge*, 140-159.

Oard, D. W., & Diekema, A. R. (1998). Cross-language information retrieval. *Annual review of Information science*, 33, 223-256.

Och, F. J. & Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation.In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 295-302.

Och, F. J. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1), 19-51. MIT Press.

Papineni, K., Roukos, S., Ward, T. & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311-318.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*.

Sánchez-Martínez, F. & Carrasco, R. C. (2011). Document translation retrieval based on statistical machine translation techniques. *Applied Artificial Intelligence*, 25(5), 329-340.

Stolcke, A. & others. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings of the international conference on spoken language processing*, 2, 901-904.

Yang, C. C. & Wing Li, K. (2004). Building parallel corpora by automatic title alignment using length-based and text-based approaches. *Information processing & management*, 40(6), 939-955.

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 4, December 2012, pp. 33-48

領域相關詞彙極性分析及文件情緒分類之研究

Domain Dependent Word Polarity Analysis for Sentiment Classification

游和正*、黃挺豪*、陳信希*

Ho-Cheng Yu, Ting-Hao (Kenneth) Huang, and Hsin-Hsi Chen

摘要

情緒分析乃近年來發展迅速之一熱門研究領域，旨在透過文本分析技術探討作者之意見傾向與情緒狀態。其中，以情緒詞與情緒詞典爲基礎之各種方法尤爲知名。然而，情緒詞之情感傾向及其行爲於不同領域文本中之行爲並不盡然相同。本研究聚焦於情緒詞彙於不同領域文本中之行爲，對房地產、旅館、和餐廳等三種不同領域之文本進行分析，並發現部分情緒詞彙於不同領域文本中的情緒傾向非但有差異，甚至彼此衝突。此外，部分未收錄於情緒詞典中之「非情緒詞」，在特定領域中亦可能成爲「領域相依」之詞彙，影響情緒分類。本研究繼而提出不同詞彙權重計算方式，將此資訊加入舊有情緒分類系統中。在使用 LIBSVM 的線性核函數方式，對房地產、旅館、和餐廳等三種語料使用 5 次交叉驗證方式進行分類。實驗結果顯示所提出之 TF-S-S-IDF 分類方法，結合 TF-IDF、臺灣大學情感詞典，及計算語料之領域極性情感傾向程度(SO)，強化領域相關及領域不相關之情緒詞之權重，通過 t 檢定有效提升各領域中文件分類之效能。

關鍵詞：文件情緒分類、詞彙極性分析、機器學習

Abstract

The researches of sentiment analysis aim at exploring the emotional state of writers. The analysis highly depends on the application domains. Analyzing sentiments of

* 國立臺灣大學資訊工程學系 Department of Computer Science and Information Engineering, National Taiwan University
E-mail: {p98922004, r96944003, hhchen}@ntu.edu.tw

the articles in different domains may have different results. In this study, we focus on corpora from three different domains in Traditional and Simplified Chinese including real estate, hotel and restaurant, then examine the polarity degrees of vocabularies in these three domains, and propose methods to capture sentiment differences. Finally, we apply the results to sentiment classification with LIBSVM (linear kernel). The experiments show that the proposed method TF-S-S-IDF which integrates TF-IDF, NTU Sentiment Dictionary, and word sentiment orientation degree in each specific domain can effectively improve the sentiment classification performance.

Keywords: Document Sentiment Classification, Word Polarity Analysis, Machine Learning.

1. 緒論

隨著資訊科技發展，如今任何人皆可在任何時間、任何地點，透過智慧型手機或電腦輕易連上網路。而網路使用者除由網上獲得所需資訊外，亦可進一步發表意見、評論議題，甚或公開抒發內心的情緒。隨著社群網站興起，此類攜帶主觀意見或情緒之文本資料量大幅提升，情緒分析技術也日益受到重視，並發展出許多重要之應用，諸如產品趨勢分析、關連性商品挖掘、推薦系統、議題偵測等等。

昔日的網站內容提供者僅具有單方向的資訊傳播行爲，但隨著 Web 2.0 概念興起，如今已漸由單向轉爲雙向的互動傳播，使用者間的相互影響因而益加顯著（高, 2012; Tang & Chen, 2011, 2012）。（Pang & Lee, 2008）整理近年之研究而歸納得出下列幾項重要現象：

1. 81%的網路使用者曾對有興趣的產品進行線上研究。
2. 20%的人每天都會查看有興趣產品的評論（Review）。
3. 網路上對餐廳、旅館、或其他服務（醫生、旅行業者）的使用者評價，有 73%到 87%的程度影響使用者是否購買的慾望。
4. 相較於評價爲四顆星的商品，消費者更願意多支付 20% - 99%去購買評價爲五顆星的商品。

這些現象也說明產品的意見評論對消費者的購買意願有著重大的影響力，讓企業可以發現其商品與服務的優劣，進行改善與強化。

由於情緒分析的結果會隨著不同領域或主題而有不同，在過去相關文獻中，中文方面的多領域之情緒分析相關研究並不多見，大都是對主題式相關之單一領域文章進行研究。本研究嘗試以不同領域之語料進行分析，探討文章領域、詞彙極性、與情緒分類三者之間之關聯，繼而達到下列兩項研究目的：（一）分析不同領域、不同語言文章中詞彙極性之程度，以了解在不同的語料中，詞彙極性變化的情形。（二）研究如何將詞彙極性之變化應用於 SVM 情緒分類器，以提高文件情緒分類之準確度。

2. 相關研究

情緒分析由文章、句子、詞彙等三個不同層面切入（Ku & Chen, 2007; Ku, Huang & Chen, 2009, 2010），由單語到多語（Seki *et al.*, 2007）。情緒分類常將輸入文本分爲「正面」與「負面」兩個極性（Turney & Littman, 2003; Turney, 2002; Chaovalit & Zhou, 2005），或分爲特定情緒，如：快樂、難過、溫馨、有趣、驚訝等類別（Sautera, Eisner, Ekman, & Scott, 2010）。情緒分類也因實際應用上之需求，亦常伴隨情緒強度標註（Lu, 2010）。諸如一極佳之商品與一尚可之商品，雖皆歸爲正面評價之範疇，但強度卻不盡相同。

過去常用之情緒分類方法，主要可以分成非監督式學習法（Unsupervised Learning）與監督式學習法（Supervised Learning）兩類。Chaovalit and Zhou（2005）比較這兩類方法，發現監督式學習法具有較高的準確率，但需花費大量時間對標記完成之語料進行訓練；而非監督式學習法的效能則仰賴其所參照的詞性標記程式（POS tagger）。該實驗結果顯示後者準確率不及監督式學習法，卻具有即時性（Real-Time）的優勢。

Tan and Zhang（2008）則對各種監督式學習法進行比較，發現支援向量機（SVM）分類效果較佳。Lan *et al.*,（2005）提出了 SVM 的 TF-RF 權重計算方式並與其他方法比較，發現該方法具有最佳的準確率 。Martineau and Finin（2009）也提出了 SVM 的 DELTA TF-IDF 權重計算方式，分類效能勝過了 SVM 的 TF-IDF 權重計算方法。

3. 語料

3.1 資料蒐集及標記

本研究採用三種不同領域的語料，分別爲房地產新聞摘要、大陸旅館評論、以及臺灣餐廳評論，這些語料分屬正體與簡體中文，除編碼不同外，亦可探討語言使用上的差異。

3.1.1 房地產新聞摘要

房地產新聞摘要爲內政部建築研究所與政治大學臺灣房地產研究中心每季公布之房地產景氣動向季報內新聞摘要（Chin, 2010），並由該房地產相關研究人員對新聞摘要進行標記，該語料由 2001 年至 2010 年共 2,389 篇，與其他語料相比，篇幅較少，表 1 列出正面和負面的範例。

表 1. 房地產新聞摘要範例

正面	加速活絡市場，不動產邁向證券化。爲加速活絡不動產市場，行政院院會訂今日通過「不動產證券化條例」草案，受益證券買賣均可免徵證交稅；信託財產爲土地者，地價稅按基本稅率千分之十計徵；不動產投資信託基金或不動產資產信託持有之土地或建築物，得依地方政府規定減免地價稅及房屋稅等。
負面	再創新高，全台貧富差距飆至 66 倍。過去 11 年來，貧富差距一路狂飆，毫無減緩的趨勢。據財稅資料中心統計，1998 年，最富有的 5%與最窮的 5%的平均所得相差 32 倍，11 年後，這項所得差距已擴大至 66 倍，非常驚人。在金融風暴的衝擊下，國內失業率大幅上揚，預估這項所得差距仍會進一步竄升。

3.1.2 旅館評論

旅館住宿心得評論爲譚松波先生搜集之中文情感挖掘語料之一[1]，以進行學術研究使用，其內容爲一般民眾住宿旅館後之心得評論，再由第三方標記者標記其情緒極性。其範例如表 2。

表 2. 旅館評論範例

正面	酒店很干净，服务员会推荐我到女士无烟层，设施也比较好，餐厅的小点心味道也可以。吸取了之前美酒店不好停车的经验，在这的三天还是我比较满意的。
負面	这个酒店，隔音太差，都是用木板隔的。有纯净饮用水口，但是坏了,也不配给免费矿泉水。房间设置也是不好的。我的同事房间淋浴口竟然不出水。奉劝大家不要去住。

3.1.3 餐廳評論

餐廳評論由愛評網[2]撈取 2006 年至 2009 年之餐廳使用者評論，從其中抽取正負面各 2,000 篇評論進行實驗，其語料文章篇幅較長，表 3 是正面及負面評論之範例。

表 3. 餐廳評論範例

正面	無意間看到牛肉麵節的評審，便迫不及待的跑來 白鐵製的煮麵台，感覺就是將一般麵店放在外面的煮麵台，搬進店裡來煮，很有意思 價目表，點麵時就要先決定辣度，那的牛肉麵是用紅燒的 大碗牛肉麵，中辣，135， 牛肉給的很多，而且就如王瑞瑤小姐所說:『採用不限部位的本地牛肉，所以吃起來有硬 有軟，富嚼勁又甘甜，』有肉，有肋條，也有筋 空心菜有點太爛了 麵條是一般的麵條，但是煮的非常好，咬起來有一股韌性，非常好 接下來喝一口湯，哇，好濃啊，吃過這麼多>牛肉麵，這是我喝過牛肉味最濃的湯了，…
負面	台--市-----路的--真的很濫，之前吃就有點感覺不好，最近聚餐，朋友剛好又找這家，感覺真的差啦， 8 人吃，點 10 人份，結果感覺端來只有 2-3 人份;再加點 10 人份，結果只來 5 片肉，服務生說吃不夠再點，這擺明有點故意啦.因爲已跟不同服務生都反應，我們桌上都空啦;烤肉烤到最後在吃火鍋，因爲沒東西烤.我覺得既然開吃到飽，至少不要怕人家點吧，不然就開單點的方式就好啦.花錢是高興的，結果吃的超不高興的.真的差勁到極點.各位網友，台--店都倒光，也不要去這吃….

[1] http://www.searchforum.org.cn/tansongbo/senti_corpus.jsp

[2] http://www.ipeen.com.tw/

表 4 整理這三份語料之特性，由不同屬性比較他們之相同和差異處。

表 4. 旅館評論範例

語料	房地產新聞摘要	旅館評論	餐廳評論
文章類型	新聞摘要	個人評論	個人評論
作者背景	記者	一般民眾	一般民眾
標記方式	由房地產專家標記	由語料製作者標記	評論者自行標記
文章內容	房地產	旅館	餐廳
平均文章長度	約 130 字	約 70 字	約 300 字
文字	正體中文	簡體中文	正體中文
文章數	正面：1,418 篇 負面：971 篇	正負各 2,000 篇 共 4000 篇	正負各 2,000 篇 共 4000 篇
期間	2001~2010	未提供	2006~2009

3.2 斷詞方法

本研究採用mmseg4j[3]進行斷詞，mmseg4j乃爲一基於詞典之斷詞方式。mmseg (Tsai, 1996) 演算法主要區分爲簡單（Simple）與複雜（Complex）兩種方式進行解析，此兩種方式都是使用最大匹配演算法（maximum matching algorithm）（Chen & Liu, 1992）進行處理，其簡單的方式準確率達 95%，而複雜方式準確率達 98%。由於mmseg4j之詞典可自行擴充，故本研究除了將臺灣大學情感詞典（NTUSD）（Ku & Chen, 2007）導入外，也將語料透過中研院CKIP[4]斷詞系統斷開之詞彙導入，如圖 1。

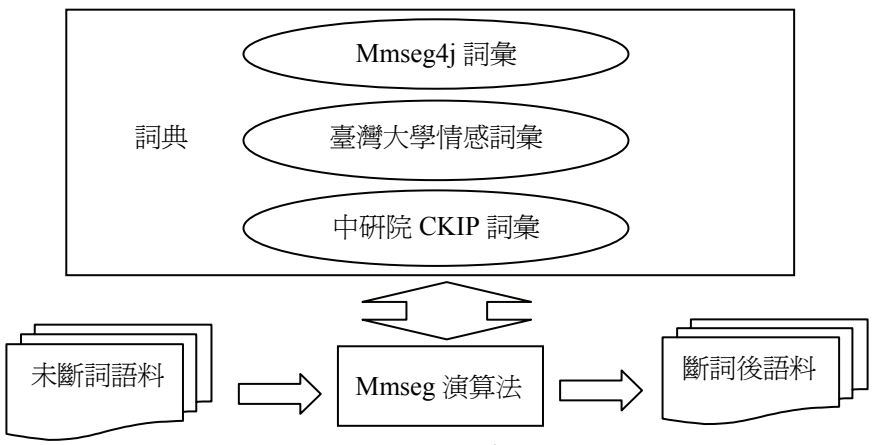


圖 1. 斷詞方法

[3] http://code.google.com/p/mmseg4j/

[4] http://ckipsvr.iis.sinica.edu.tw/

表 5 是對照擴充前後斷詞結果的比較，擴充後有效地將領域特殊詞彙斷出，例如：地價稅、房屋稅等專有名詞，故比未擴充前詞庫之斷詞效果更佳。

表 5. 詞庫擴充前後斷詞比較（差異處以粗體表示）

原文	爲加速活絡不動產市場，行政院院會訂今日通過「不動產證券化條例」草案，受益證券買賣均可免徵證交稅；信託財產爲土地者，地價稅按基本稅率千分之十計徵；不動產投資信託基金或不動產資產信託持有之土地或建築物，得依地方政府規定減免地價稅及房屋稅等。
擴充前斷詞結果	[爲, 加速, 活絡, 不動產, 市場, 行政院, 院會, 訂, 今日, 通過, 不動產, **證券, 化,** 條例, 草案, 受益, 證券, 買賣, 均可, 免, 徵, **證, 交稅,** 信託, 財產, 爲, 土地, 者, **地價, 稅,** 按, 基本, 稅率, **千分之, 十, 計, 徵,** 不動產, 投資, 信託, 基金, **或不, 動產,** 資產, 信託, 持有, 之, 土地, 或, 建築物, **得, 依,** 地方政府, 規定, 減免, **地價, 稅,** 及, **房屋, 稅,** 等]
擴充後斷詞結果	[爲, 加速, 活絡, 不動產, 市場, 行政院, 院會, 訂, 今日, 通過, 不動產, **證券化,** 條例, 草案, 受益, 證券, 買賣, 均可, 免, 徵, **證交稅,** 信託, 財產, 爲, 土地, 者, **地價稅,** 按, 基本, 稅率, **千分之十, 計徵,** 不動產, 投資, 信託, 基金, **或, 不動產,** 資產, 信託, 持有, 之, 土地, 或, 建築物, **得依,** 地方政府, 規定, 減免, **地價稅,** 及, **房屋稅,** 等]

4. 分析

4.1 情緒度程度計算方式

本研究以情緒詞彙於情緒語料庫中之正負面文章頻率（Document Frequency）來分辨該詞彙之極性程度，亦即當一詞彙在較多正面文章中出現、在較少負面文章中出現時，此一詞彙便較偏向正面；反之則越偏向反面。利用此特性，吾人可進而計算文章中詞彙之情感傾向程度（Sentiment Orientation），並區別文章的正負面極性。其計算方式如公式(1)、(2)、及(3)所示。

$$PSO = \frac{|D_{w\ in\ P}| + 1}{|D_P|} \tag{1}$$

$$\mathrm{N}SO = \frac{|D_{w\ in\ N}| + 1}{|D_N|} \tag{2}$$

$$SO = \log_e\left(\frac{PSO}{NSO}\right) \tag{3}$$

PSO 爲正面情感傾向程度（Positive Sentiment Orientation），是由詞彙所在的正面文章數量$|D_{w\ in\ P}|$除以所有正面文章數量$|D_P|$。爲了避免該詞彙在某類文章中未出現，我們將$|D_{w\ in\ P}|$加 1，以避免 PSO 爲 0。同理，NSO 爲負面情感傾向程度（Negative Sentiment Orientation），是由詞彙所在的負面文章數量$|D_{w\ in\ P}|+1$除以所有負面文章數量$|D_P|$。

得到 PSO 與 NSO 後進行比較，如 PSO 大於 NSO 則情感傾向爲正面，反之情感傾向爲負面。當 $\frac{PSO}{NSO}$ 等於 1 時，代表 PSO = NSO，應爲無情緒程度，故 $\log_e(\frac{PSO}{NSO})$ 即爲 0。本研究將 $\log_e(\frac{PSO}{NSO})$ 定義爲情感傾向程度（Sentiment Orientation，SO），其絕對值代表情感強烈程度，而正負號代表正負向。

4.2 詞彙分析

在表 6、表 7、和表 8 中，我們分別由房地產、旅館、和餐廳語料中，各抽取出 SO 最正面及最負面之五個詞彙，進行分析和討論。詞彙後方的「+」符號，代表情感詞典裡之正面詞彙，「-」符號則爲情感詞典裡之負面詞彙。

除特定情感詞彙外，還有許多領域相關的特定詞彙，例如：「搶標」、「證券化」、「重劃區」等皆在其中。旅館評論之正負面詞彙較多爲情緒詞典內的詞彙，吾人可注意到詞彙「美中不足」似乎是完美、但仍有缺陷的意思，以句子而言，可能有負面的傾向，但若探討整篇文章的情緒傾向，考慮語境時則明顯爲一個正面的詞彙。另外，可觀察到餐廳評論之情緒傾向正面之詞彙大都爲餐廳名稱或是食品名稱，而情緒辭彙較少，推測應爲當一間餐廳或是一項食品，當較常出現於正面文章內，其名稱就會使得文章偏向正面，例如，他做的蛋糕和阿默一樣，而阿默經常出現在正面文章，使得阿默這個辭彙就有讚揚的正面情緒。類似這類型的詞彙是依照語料的特性所賦與詞彙的極性，可以稱作爲領域相關的情緒詞彙，依照不同的語料可能會有不同的極性，反觀傾向負面之詞彙則較多情緒詞典詞彙。

表 6. 房地產正負極性詞彙

SO 最高之 5 個詞彙				SO 最低之 5 個詞彙			
詞彙	SO	$\lvert D_{w\ in\ P}\rvert$	$\lvert D_{w\ in\ N}\rvert$	詞彙	SO	$\lvert D_{w\ in\ P}\rvert$	$\lvert D_{w\ in\ N}\rvert$
搶標	2.99	28	0	警訊-	-3.64	0	25
發行+	2.82	25	0	下修	-3.39	2	60
證券化	2.64	43	1	跌幅	-3.37	1	39
新高價	2.54	19	0	不吃不喝	-3.21	0	16
商機+	2.49	18	0	慘	-3.09	0	14

表 7. 旅館評論正負極性詞彙

SO 最高之 5 個詞彙				SO 最低之 5 個詞彙			
詞彙	SO	$\lvert D_{w\ in\ P} \rvert$	$\lvert D_{w\ in\ N} \rvert$	詞彙	SO	$\lvert D_{w\ in\ P} \rvert$	$\lvert D_{w\ in\ N} \rvert$
一流+	3.47	31	0	爛-	-3.99	0	53
美中不足	3.00	19	0	最差-	-3.95	1	103
略顯	3.00	19	0	要我	-3.33	0	27
相當不錯+	2.94	18	0	惡劣-	-3.26	1	51
首選+	2.94	18	0	折騰-	-3.22	0	24

表 8. 餐廳評論正負極性詞彙

SO 最高之 5 個詞彙				SO 最低之 5 個詞彙			
詞彙	SO	$\lvert D_{w\ in\ P} \rvert$	$\lvert D_{w\ in\ N} \rvert$	詞彙	SO	$\lvert D_{w\ in\ P} \rvert$	$\lvert D_{w\ in\ N} \rvert$
饗食	3.22	24	0	不耐煩-	-3.14	0	22
竹笙	3.09	21	0	有待加強-	-3.04	0	20
阿默	3.04	20	0	無言	-2.97	1	38
荷蘭	3.00	19	0	臉色	-2.94	0	18
京兆尹	3.00	19	0	老街	-2.83	0	16

表 9. 極性衝突的詞彙

詞彙	房地產	旅館	餐廳	詞彙	房地產	旅館	餐廳
才能+	-2.32	-1.45	0.13	很快+	-1.48	2.04	0.31
不夠好-	0.00	1.10	-2.40	按照+	0.00	-1.53	2.08
不會有	0.00	-2.20	0.71	恐怕-	-2.58	-0.92	0.16
及時+	-1.48	1.12	1.61	酒鬼-	0.00	-1.10	2.08
少數-	1.76	1.10	1.01	高檔	-1.29	1.10	1.53

進一步對這三種語料之詞彙極性交叉比對後發現，某些詞彙在某一語料傾向正面，但在其他語料卻傾向負面，或是在某語料傾向中性，但在其他語料卻有正負傾向。部分具有衝突的詞彙，如表 9 所示。以表 9 之「恐怕」一詞爲例，於房地產新聞及旅館評論中，該詞多爲負面意思；然而於餐廳評論中，卻多用於「恐怕得排隊」之意，故多出現於正面文章。由此可知，部分詞彙儘管並無強烈「領域相關性（domain relatedness）」——如「恐怕」一詞與房地產、旅館或餐廳皆無明顯關聯——卻仍會因領域的不同而產

生不同的正負情緒傾向，即具有「領域相依性（domain dependency）」。

以下表10 和表 11 以「不夠好」與「遙遙領先」爲例，進一步說明此現象。「不夠好」雖爲負面的情緒詞彙，但在旅館評論中表現出「盼望未來能夠更好」的正面情緒，故文章多傾向正面；但在餐廳評論中則多有責備之意，故情緒傾向負面。而「遙遙領先」雖爲正面詞彙，在餐廳評論裡亦多用於形容環境與美味；但在房地產新聞中，該詞多用來形容房價、油價及其他財經指標高漲，因此多用於負面文章。

表 10. 衝突詞彙範例「不夠好」（以粗體標示）

旅館正面評論	房间很清洁，洗手间装修好。早餐非常丰盛，豆腐脑很好吃！交通很方便。隔音**不够好**，我们的邻居夜里 1 点多电视声音开的很大，投诉后酒店服务人员去与隔壁邻居协商解决了。
	房间还算干净整洁,服务也可以,以这个价格来说,不错了,建议大家要定有窗户的房间,但是餐厅**不够好**,有一次吃的蘑菇是酸的,附近没有超市,不太方便,买水果都要走很远,
餐廳負面評論	Ｙ勒勒～可能是我好料吃多了！ 所以這家就相形失色了吧！ 或著是這家分店不比老店好吃！ 品管**不夠好**嚕～ @0@ 套餐所附的小菜！ 我也來介紹一下吧～ 都是一碟有兩種小菜！….可是是這家分店完全沒有提供！ 或是告知是否有這些好康・・・ 真的是弱掉了～～ ...
	用餐時間從 3 點到 5 點 送完餐之後沒有再來加過一次水 也沒有任何人來收吃完後的盤子 價格不美麗，東西不好吃，服務又**不夠好** 這樣的餐廳，如何讓人再想踏進一步.....

表 11. 衝突詞彙範例「遙遙領先」（以粗體標示）

餐廳正面評論	這雪片冰果真名不虛傳的啊！好吃的程度真的是**遙遙領先**，尤其是咖啡口味的呢！ 巧克力口味有點讓我心痛就是了，下再來吃要點抹茶紅豆啊！炎炎夏日吃冰果真很透心涼耶 咖啡口味：★★★★★ 巧克力口味：★★★ 炎炎夏日一定要好好享受吃冰的喜悅阿！
	嚴格說起來 到今天造訪野之前 在臺灣還沒吃過令我驚豔的螢烏賊 猶如新鮮荔枝般的滑脆口感 加上肚子裡的爆漿海味 備材選料的水平果然**遙遙領先**能用螢烏賊捏製握壽司 可見深厚功底 玉子 口感豐厚扎實 混入滿滿磨碎白蝦 明顯的海洋鮮香 不輸給築地的專業玉子舖
房地產負面新聞	高學歷失業惡化，創 23 年新高。國內勞動市場出現警訊！今年八月失業率爲四・一四％，創下三年來新高；高學歷、高失業率問題更加速惡化，八月大學以上失業率五・二六％，不僅**遙遙領先**其他教育程度，更創下二十三年來新高
	2010 年第 1 季全台新屋成交價突破歷史新高，北市以每坪 56.6 萬元**遙遙領先**其他城市，但銷售率小幅萎縮，議價空間拉大.17.16％。學者擔憂，房價反彈無基本面支撐，且市場觀望氣氛逐漸形成，V 型反彈極可能會轉成 M 型修正，最快今年下半年到明年就會看到。

5. 研究方法

5.1 實驗設定

透過前述的研究，本研究提出以下三種不同類型之的特徵值進行分類：一，使用單一詞彙（Unigram）。二，使用臺灣大學情感辭典（NTUSD）之詞彙資訊。三，使用二元詞彙（Bigram）。

在以下實驗中，我們將比較這三種方法的優劣。而特徵值則兼採詞頻（Term Frequency）與情感傾向程度（SO），詳細計算方式如表12。表12 第一直行爲方法代碼，第二行是計算方式，第三行是公式中所使用的符號說明。

表 12. 權重計算方式

方法	計算方式	說明
TF-SO	$TFSO = TF \times (\lvert SO \rvert + 1)$	TF：詞彙在該文章出現的數量除以該文章的詞彙總數；|SO|：情感強烈程度[5]，數值越大，代表情感程度越強烈時，也越爲重要。在以SVM分類時，當|SO|爲0，則維持TF之權重，因此將|SO|+1。
TF-SO-IDF	$TFSOIDF = TFIDF \times (\lvert SO \rvert + 1)$	這是一種結合|SO|與 TF-IDF 的方法，使得 IDF 與 SO 相互影響。若該詞彙情緒程度越強，則 TF-IDF 越重要，實驗中並與領域無關之情感詞典比較。
TF-SD-IDF	$TFSDIDF = \begin{cases} TFIDF \times k & if\ w\ in\ \text{NTUSD} \\ TFIDF & \text{otherwise} \end{cases}$	結合 TF-IDF 和情感辭典，若該詞彙在情感詞典裡出現，代表越重要，則進行加權(乘上 k)。由於搭配情感辭典，因此在 Bigram 分類並不使用。
TF-S-S-IDF	$TFSSIDF = \begin{cases} TFSOIDF \times k & if\ w\ in\ \text{NTUSD} \\ TFSOIDF & \text{otherwise} \end{cases}$	結合 TF-SO-IDF 和情感辭典，若該詞彙在情感詞典裡出現，代表越重要，則進行加權(乘上 k)。由於搭配情感辭典，因此在 Bigram 分類並不使用。

TF-SD-IDF及TF-SS-IDF皆須搭配加權之k值，而k值之決定，本研究以房地產評論，採用不同的k值來進行分類，並觀察最後的準確率以決定k值，分類的結果如表13。

表 13. 房地產語料不同 k 值分類結果

k	1.0	1.5	**2.0**	3.0	5.0	10.0
Accuracy	0.848	0.851	**0.852**	0.840	0.816	0.7972

[5] 線性 SVM 分類方式在尋找參數過程會自動將正負號進行調整，故取絕對值差異影響不大，這裡是以情緒強烈程度來對應 TF-IDF，故取絕對值較容易理解。

由上表可以發現，k 值在 2.0 時所得到的準確率最佳，而在後續的 3.0 與 5.0 及 10.0 皆不斷遞減，故本研究後續之 k 值皆以 k = 2.0 來進行分類。

實驗結果並與TF-IDF、TF-RF（Lan, Sung, Low, & Tan, 2005）、及DELTA TF-IDF（Martineau & Finin, 2009）等三種方式進行比較，這三種方法的說明如表14。

表 14. 基準方法權重計算方式

基準方法	計算方式	說明
TF-IDF	$TFIDF = TF \times log_e(\frac{\|D\|}{DF})$	TF：爲詞彙在該文章出現的數量除以該文章的詞彙總數；$\|D\|$：文章的總數；DF：含有該詞彙的文章數量。
TF-RF	$TFRF = C_{t,d} \times \log_e(\frac{\|D_{w\ in\ P}\|}{\|D_{w\ in\ N}\|} + e)$	$C_{t,d}$ 爲詞彙在該文章出現的數量，$\|D_{w\ in\ P}\|$ 爲詞彙所在的正面文章數量，$\|D_{w\ in\ N}\|$ 爲詞彙所在的負面文章數量，兩者相除後，加上 e。原文以 2 爲底，故加 2，但本研究之 log 皆以 e 爲底，故修改成加 e。
Delta TF-IDF	$Delta\ TFIDF = C_{t,d} * \log_e(\frac{\|D_P\| \times \|D_{w\ in\ N}\|}{\|D_{w\ in\ P}\| \times \|D_N\|})$	$C_{t,d}$ 爲詞彙在該文章出現的數量，$\|D_{w\ in\ P}\|$ 爲詞彙所在的正面文章數量，$\|D_{w\ in\ N}\|$ 爲詞彙所在的負面文章數量，$\|D_P\|$ 爲正面文章數量，$\|D_N\|$ 爲負面文章數量。

透過上述方法，計算出權重後，再使用LIBSVM[6]的線性核函數（Linear Kernel Function）方式進行分類。由於線性方式僅需單一參數（Cost），不但速度優於其他核函數，文件分類效能也不會輸於其他核函數（Lan, Sung, Low, & Tan, 2005），故選擇此一方式。本研究將依序嘗試不同的參數（Cost），以最佳的結果爲主。

實驗中將三種語料使用 5 次交叉驗證（5-Fold Cross Validation）方式進行，即計算權重時(包括上述的所有權重值)僅使用訓練資料集來計算，以避免內部測試(Inside Test) 的問題。同時，本研究使用準確度（Accuracy）以評估效能，並 t 檢定進行有效性檢定。

6. 實驗結果

6.1 單一詞彙（Unigram）分類結果

表15 是採用單一詞彙T檢定分類結果，評估的標準是準確率，對TF-S-S-IDF與其他方式進行比較，若顯著程度達到 95%的信心水準則標記爲*，若達到 99%則標記爲**，而達到 99.5%則標記爲***。吾人發現僅使用單一之TF-SO或TF-IDF的效果很接近，但是將IDF

[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

與SO相乘，也就是TF-SO-IDF，其效果更好，TF-SO-IDF優於其他兩種。若結合情感辭典，可將分類效果更進一步提昇。表中呈現TF-S-S-IDF優於TF-SO-IDF，TF-SD-IDF優於TF-IDF。總結，Unigram的結果以TF-S-S-IDF爲最佳，TF-SO-IDF與TF-SD-IDF次之，接著是TF-IDF，與其他方法。這種情況以餐廳的語料最爲顯著。

表 15. Unigram 分類 T 檢定綜合比較表(Accuracy)

語料	TF-IDF	TF-RF	Delta	TF-SO	**TF-SO-IDF**	TF-SD-IDF	**TF-S-S-IDF**
房地產	0.848*	0.849	0.853	0.847*	0.854	0.852	**0.863**
旅館	0.916	0.906***	0.914*	0.915	**0.924**	0.918	0.923
餐廳	0.861**	0.839***	0.849***	0.854***	0.871	0.869	**0.875**

6.2 僅用情緒詞典詞彙分類結果

表16 顯示僅使用情緒詞彙來進行分類，同樣的，也對TF-SO-IDF與其他方法進行T檢定，標記方式同表15。與前一方式分類結果比較，各方法之準確度皆顯著下降。這樣的結果顯示，不在情緒辭典內的領域相關情緒辭彙也在分類器中扮演著重要的角色，因此造成此分類結果效能顯著下降。另外，可以觀察到僅用情緒詞的結果並不一致，在房地產分類中，以Delta最佳，在旅館分類中則是以TF-SO-IDF最佳，而在餐廳的評論中則是以TF-IDF最佳。

表 16. Unigram 僅用情緒詞分類綜合比較表(Accuracy)

方法	TF-IDF	TF-RF	Delta	TF-SO	**TF-SO-IDF**
房地產	0.784	0.786	**0.796**	0.784	0.787
旅館	0.883	0.877*	0.878	0.880	**0.887**
餐廳	**0.806**	0.795	0.792	0.783	0.804

6.3 二元詞彙（Bigram）分類結果

表17 是使用二元詞彙（Bigram）的分類結果，同樣的，也對TF-SO-IDF與其他方法進行T檢定，標記方式如前所述，與僅用情緒詞彙分類相比，除餐廳語料外，其他語料的分類效能又下降些許。而在房地產是以TF-IDF最佳，其他爲TF-SO-IDF。

表 17. Bigram 分類綜合比較表(Accuracy)

語料	TF-IDF	TF-RF	Delta	TF-SO	**TF-SO-IDF**
房地產	**0.777**	0.769	0.756	0.767	0.767
旅館	0.879	0.858***	0.879	0.883	**0.886**
餐廳	0.805	0.778***	0.766***	0.793	**0.811**

7. 結論

本研究以文章頻率來計算詞彙極性程度，探討了各領域詞彙的差異，應用詞彙極性程度結合情緒辭典進行情緒分類，得到更佳的分類效果。從實驗過程的分析與討論，可以歸納以下幾點結論：

1. 情緒詞彙可以歸類爲兩類，即「領域不相關」之詞彙與「領域相關」之詞彙。「領域不相關」之詞彙主要爲語言學上帶有情感之詞彙，大都爲情緒詞典內之詞彙。而「領域相關」之詞彙則是透過語料訓練的方式取得，依照各領域不同而有不同。例如，京兆伊、鼎泰豐、捷運、重劃區等詞彙，原本並未帶有情緒程度，但從訓練的語料得出爲正面之情緒詞彙，即屬於此類。這類「領域相關」之情緒詞彙，若用於其他領域則可能不成立，即「鼎泰豐」若用於房地產及旅館則非此類情緒詞彙。

   「領域相關」之情緒詞彙在具有對立關係之事物尤爲顯著，舉凡在宗教上或在政治上甚至是社會上，都可見到，舉例而言：某政治人物之姓名，若出現在其政治立場相左之報紙社論之語料內，則該姓名雖非爲情緒詞，但是極有可能成爲本研究之「領域相關情緒詞彙」。這類由原本的「非情緒詞彙」，轉變成「領域相關情緒詞彙」，在未來甚至也有可能成爲各領域共通之情緒詞，這類的轉換，本研究稱之爲「詞彙情感注入」現象， 例如：三國時「諸葛亮」與「阿斗」原本也僅僅是人物名稱，但透過三國歷史的語料不斷習得，皆使得「諸葛亮」與「阿斗」成爲情緒詞彙。

2. 不同領域間詞彙情緒程度可能會有衝突，可歸因不同語料所用之匹配詞彙不同，造成差異。例如，「遙遙領先」在房地產語料內與失業率及高房價搭配，造成情緒極性相反，故此類詞彙雖在情緒詞典，但其仍可歸類爲領域相關之情緒詞彙。

3. 若比較「Unigram 所有詞彙」與「Unigram 僅用情緒詞彙」之結果，發現語料內若領域相關情緒詞彙較多時（房地產與餐廳語料），僅用情緒詞典之分類效果較差。反之，若語料內之領域相關情緒詞彙較少時（旅館語料），僅用情緒詞典之分類效果較好。但橫向比較「Unigram 所有詞彙」之 TF-IDF 與 TF-SD-IDF（TF-IDF 情緒辭典加權）以及 TF-SO-IDF 與 TF-S-S-IDF（TF-SO-IDF 情緒辭典加權）卻可以發現，情緒辭典作用在房地產與餐廳語料的效果更佳。

4. 透過詞彙 SO 的計算，找出領域相關之情緒詞彙之極性程度，加入 TF-IDF，再搭配 SVM 分類方法，與僅用 IDF 或僅用 SO 相比之下，兩者相互的結合，更可以提昇分類效能。

5. 本研究提出之 TF-S-S-IDF 分類方法，結合 TF-IDF、臺灣大學情感辭典，及計算語料之領域極性 SO，強化領域相關及領域不相關之情緒詞之權重，得出更佳的分類效能。

6. 中文在預處理上的斷詞結果不同，也將會影響詞彙所內含的意義，能斷出越長的詞彙，並不代表斷詞效果越佳。因爲較長的詞彙出現的頻率越少，造成資訊的不全。例如，「房間的設施」與「房間設施」，若皆斷成一個詞彙則會造成不同。但是相對的，若是所斷出的詞彙越短（單一字），則會造成斷出的詞彙僅具字面意義而造成誤判。

致謝

Research of this paper was partially supported by National Science Council (Taiwan) under the contract NSC 98-2221-E-002-175-MY3 and 2012 Google Research Award.

參考文獻

Chaovalit, P. & Zhou, L. (2005). Movie review mining: a comparison between supervised and unsupervised. In *Proceedings of the 38th Hawaii International Conference on System Sciences*.

Chin, Y.-L. (2010) A review and discussion of real estate cycle indicators analysis and publication method. Research Project Report. Architecture and Building Research Institute, Ministry of the Interior, Taiwan. http://www.abri.gov.tw/

Ku, L.-W. & Chen, H.-H. (2007). Mining opinions from the web: beyond relevance retrieval. *Journal of American Society for Information Science and Technology*, 58(12), 1838-850.

Ku, L.-W., Huang, T.-H., & Chen, H.-H. (2009). Using Morphological and Syntactic Structures for Chinese Opinion Analysis. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 1260-269.

Ku, L.-W., Huang, T.-H., & Chen, H.-H. (2010). Construction of Chinese Opinion Treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 1315-319.

Lan, M., Sung, S.-Y., Low, H.-B., & Tan, C.-L. (2005). A comparative study on term weighting schemes for text categorization. In *Proceedings of International Joint Conference on Neural Networks*, 1032-033.

Lu, Y., Kong, X., Quan, X., Liu, W., & Xu, Y. (2010). Exploring the sentiment strength of user reviews. In *Proceedings of 11th International Conference on Web-Age Information Management*, 471-482.

Martineau, J. & Finin, T. (2009) Delta TFIDF: An improved feature space for sentiment analysis. In *Proceedings of the Third AAAI International Conference on Weblogs and Social Media*, 258-261.

Pang, Bo & Lee, Lillian (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2) 1-135

Sautera, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. In *Proceedings of the National Academy of Sciences*, 107(6), 2010.

Tan, S. & Zhang, J. (2008). An empirical study of sentiment analysis for Chinese documents. *Expert System with Applications*, 34(4), 2622-2629.

Tang, Y.-J. & Chen, H.-H. (2011). Emotion modeling from writer/reader perspectives using a Microblog dataset. In *Proceedings of IJCNLP Workshop on Sentiment Analysis where AI Meets Psychology*, 11-19.

Tang, Y.-J. & Chen, H.-H. (2012). Mining sentiment words from microblogs for predicting writer-reader emotion transition. In *Proceedings of 8th International Conference on Language Resources and Evaluation*, 1226-1229.

Tsai, C.-H. (1996). MMSEG: A word identification system for Mandarin Chinese text based on two variants of the maximum matching algorithm. Available at http://technology.chtsai.org/mmseg/.

Turney, P. D. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 417-424.

Turney, P. D. & Littman, M. L. (2003). Measuring praise and criticism: inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21, 315-346.

Seki, Y., Evans, D.K., Ku, L.-W., Chen, H.-H., Kando, N., & Lin, C.-Y. (2007). Overview of Opinion Analysis Pilot Task at NTCIR-6. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, 265-278.

高宏宇 (2012)。文字探勘於社群網路研究之發揮。NCP Newsletter, 38, 11-16.

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 4, December 2012, pp. 49-68

利用機器學習於中文法律文件之標記、案件分類及量刑預測

Exploiting Machine Learning Models for Chinese Legal Documents Labeling, Case Classification, and Sentencing Prediction

林琬真*、郭宗廷*、張桐嘉+、顏厥安+、陳昭如+、林守德*

Wan-Chen Lin, Tsung-Ting Kuo, Tung-Jia Chang,

Chueh-An Yen, Chao-Ju Chen and Shou-de Lin

摘要

人工智慧於法學領域所發展出的法學資訊系統，主要提供參考資訊以協助司法審判；其重要議題包括法律文件分類、法律文件摘要、類似過去案例搜尋、協助判刑等。本論文著重探討「強盜罪」與「恐嚇取財罪」的分類，以及此兩種罪的量刑預測。我們針對「強盜罪」與「恐嚇取財罪」來定義 21 種法律要素標籤，並嘗試自動擷取所定義之標籤。實驗說明利用我們定義之法律要素標籤，能確實改善案件分類，以及進行量刑預測。最後，我們亦針對實驗結果，討論「強盜罪」與「恐嚇取財罪」的特徵，以及影響判刑長短之因素。

關鍵字：法律文件分類、量刑預測、強盜罪、恐嚇取財罪

* 國立台灣大學資訊工程系 Department of Computer Science and Information Engineering, National Taiwan University
E-mail: myownstuff@hotmail.com; d97944007@csie.ntu.edu.tw; sdlin@csie.ntu.edu.tw
+ 國立台灣大學法律學院 College of Law, National Taiwan University
E-mail: d97a21003@ntu.edu.tw; filawsof@ntu.edu.tw; tanchiauju@ntu.edu.tw

Abstract

This paper exploits machine learning methods to separate robbery and intimidation cases, and predicting their sentencing by considering defined legal factors. We introduce a framework to fetch 21 legal factor labels of robbery and intimidation cases, then use the labels for case classification and sentencing prediction. Our experiments show that the legal factor labels can indeed improve the results of case classification and sentencing prediction. We then discuss the influence of these legal factors in both case classification and sentencing prediction tasks.

Keywords: Case Classification, Sentencing Prediction, Robbery, Intimidation

1. 緒論

在審判法律案件的法律體系大致可分爲成文法(civil law)與判例法(common law)。判例法的特色在於遵循先例，法院在判決案件時依循相關判例來對目前審理的案件做判斷；成文法的特點在於有完整法典，法院依據成文法規作出判決，而判例做爲參考不會拘束日後的判決。

台灣的法律體系屬於成文法，其中刑法明訂哪些不法行爲屬犯罪行爲，以及對於這些犯罪行爲應如何科處刑罰；但在法官判決，先前之判例仍具有一定的參考價值。然而，儘管法律條文已明確列出各種犯罪行爲，在於實際判斷上仍具有模糊地帶。例如刑法中「強盜罪」與「恐嚇取財罪」具有類似不法構成要素：刑法第 328 條第 1 項定義普通強盜罪：「意圖爲自己或第三人不法之所有，以強暴、脅迫、藥劑、催眠術或他法，至使不能抗拒，而取他人之物或使其交付者，爲強盜罪，處五年以上有期徒刑」；刑法第 346 條第 1 項定義恐嚇取財罪：「意圖爲自己或第三人不法之所有，以恐嚇使人將本人或第三人之物交付者，處六月以上五年以下有期徒刑，得併科一千元以下罰金」。兩者之差異主要在於強盜罪的行爲人犯罪行爲脅迫程度，足以使得被害人不能抗拒。然而此差異在實際案例上，卻容易造成判斷混淆，例如「行爲人持槍進入超商叫被害人把錢拿出來」以及「行爲人持美工刀進入超商叫被害人把錢拿出來」，前者判屬強盜罪而後者則屬恐嚇取財罪。另一方面，兩罪的刑期相差甚大（前者是五年以上，後者是五年以下），對於嫌疑人而言影響亦甚鉅。因此，一個能支援及協助法官判別「易有模糊地帶之相關罪行」，乃至進一步提供建議刑期的系統，便極爲重要。

但是，要建置這樣的系統，會面臨數個挑戰。首先，此系統需能自動標記法律要素標籤，而不需額外之人工參與。其次，此系統需以法律要素標籤自動進行案件分類及量刑預測。最後，系統所做出的建議及預測結果，亦需經由仔細的檢查和討論，以確保其可信度。

爲了解決上述的問題，本研究針對「強盜罪」與「恐嚇取財罪」定義法律要素，並期望達成自動標記法律要素，接著利用法律要素資訊來分類「強盜罪」與「恐嚇取財罪」以及預測此兩種罪的判處刑期，最後討論「強盜罪」與「恐嚇取財罪」的分類特徵以及

影響判刑的因素。

本文之內容安排如下。在第二節，我們將探討及比較相關文獻。在第三節，我們將介紹自動標記方法及實驗結果。在第四節及第五節，我們以人工標記分別說明案件分類及量刑預測之方法和結果。在第六節，我們會結合自動標記來進行案件分類及量刑預測。最後在第七節，我們將提出結論及未來之可能方向。

2. 相關研究

人工智慧結合法律領域的法資訊學在國外已研究多年，目的在於增進司法審判的效率以及協助撰寫法律文件，主要研究包括法律文件分類、法律文件總結、相關案件檢索、協助判決等。

在於法律文件分類研究上，Ashley (Bru¨ninghaus, 2009)等人提出 SMILE 系統處理營業秘密法案件。依據事先定義之要素，以人工標記方式建立案例資料庫，並使用決策樹演算法來分類案件。Lame (Lame, 2001)等人針對法國法律文件是先建立常用字表，以 Term frequency-inverse document frequency (TF-IDF)爲權重並使用 Support vector machine (SVM)進行法律文件分類。在處理中文法律文件分類研究中，Liu (LIU, 2004)等人建立一個類 Case based reasoning (CBR)系統來分類 12 種罪，此系統依據事先定義的各種犯罪規則建立案例資料庫，並使用 k-nearest neighbor algorithm (kNN)選出與處理案件最相似之案例以達分類目的。在我們的研究中，除利用詞彙的 TF-IDF 資訊外，尚針對要分類的犯罪設計法律要素標籤以提供犯罪資訊，希望能利用 TF-IDF 以外的資訊幫助分類案件以及預測量刑。

在量刑預測研究上，Schild (SCHILD, 1998)提出一個 CBR 系統來預測強盜案件以及強制性交案件的量刑，系統詢問使用者事先定義之相關問題，根據使用者的回答利用決策樹演算法找到類似案例，最後依據過往案例來預測量刑。在處理中文法律案件量刑方面，司法院於 2011 年提出量刑資訊系統(Kuo *et al.*, 2006)，針對妨害性自主罪設計量刑因子，對於量刑因子資訊進行統計分析，提供妨害性自主罪整體量刑分布、分析量刑加重或酌減之原因。

而我們研究中則是提出的法律要素標籤資訊，並且希望以自動標記來取代人工標記。在於連續標記方面 Wei Jiang 等人(Jiang, 2006)比較 HMM、MEMM、CRF 以及 SVM 在於連續標記詞性上的效果，實驗結果顯示 CRF 在於連續標記任務上較其他模型有良好表現。另外本研究也參考 Zhang Chengmin 等人(Zhang, 2008)探討 CRF 在於標記任務上使用的特徵。

3. 自動標記法律要素標籤

每種犯罪行爲有各自的構成要素包括行爲主體、行爲客體、行爲、行爲時之特別情狀等等，不法行爲必須符合特定的「構成要件要素」，並且成立「違法性」與「有責性」才屬犯罪行爲，而本研究中單純就「構成要件要素」來對法律案件做分類及量刑預測。我

們針對「強盜罪」與「恐嚇取財罪」定義 21 種法律要素標籤，期望能表達法律案件中的犯罪構成要素。標籤依性質分爲兩大類：

- Global labels：法官認定事實、量刑、檢察官起訴內容、被告辯護、證人證詞，共 5 種。
- Local labels：行爲人、犯罪行爲、被害人、被害人反應、財物轉移態樣、行爲人主觀要素、犯案時間、犯案工具、犯案地點、行爲人特徵、被害人特徵、犯後態度、行爲人動機、行爲人與被害人關係、財物類型、共犯，共 16 種。

我們由植根法律網(植根法律網)收集 2006 年到 2010 年總共 21 個地方法院的刑事判決書，去除裁定、簡易判決、上訴駁回、不起訴、無罪等案件後，總計強盜與恐嚇取財案件共 2113 件。考量到人工標記所需之時間及人力成本，我們隨機選取了 140 件進行人工標記，案件分布如表 1，並使用中研院斷詞系統(Hotho *et al.*, 2003)對語料進行斷詞。

表 1. 人工標記判決書分布情形

地方法院	台北	士林	板橋	宜蘭	基隆	桃園	新竹	彰化	南投	台南	花蓮	台東	屏東	總件數
恐嚇取財	0	2	17	11	1	6	9	3	2	9	4	3	0	67
強盜	1	5	13	6	1	4	4	6	7	8	7	2	9	73

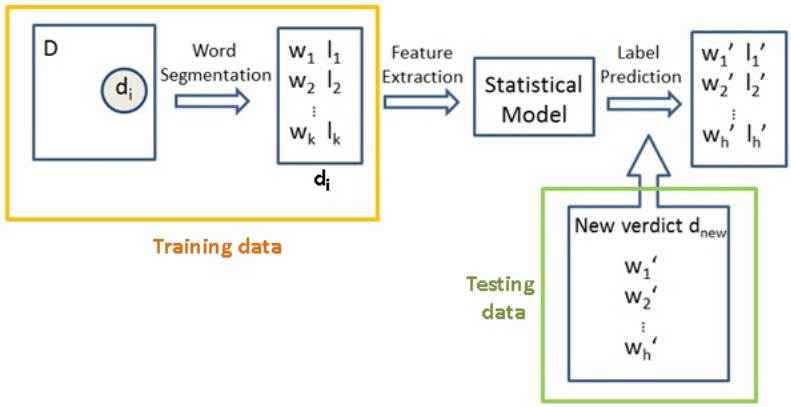


圖 1. 自動標記架構

我們定義 D 爲判決書之集合 $D=\{d_i \mid i=1,2\ldots m\}$，L 爲定義標籤集合 $L=\{l_j=1,2\ldots n, n=number\ of\ labels\}$，每篇判決書 d_i 經過斷詞處理後，得到 W_i 爲判決書 d_i 之斷詞集合，其中 $W_i=\{w_k \mid k\in N\}$。而自動標記問題定義如下：給定一篇斷詞形式的判決書 $d_{new}=W_{new}$，判斷 W_{new} 中的每個斷詞 w_k 所應標記的標籤 l_j，自動標記架構如圖 1 所示。我們運用機率模型來進行自動標記，在給定 x 爲一含有 n 個字的序列，y 爲

對應之標籤序列，機率模型中的 Conditional random fields (CRF) 模型表示式如下：

$$p(y \mid x) = \frac{exp\sum_j \lambda_j F_j(x,y)}{Z(x,\lambda)}$$

$$Z(x,\lambda) = \sum_y exp\sum_j \lambda_j F_j(x,y)$$

$F_j(x,y) = \sum_i f_j(y_{i\text{-}1}, y_i, x, i)$ ，其中 i 表在序列 x 中的位置

由於 CRF 模型中特徵函式能考量位於序列的位置，因此能考慮長距離上下文資訊，在於連續標記有極好的效能，故使用 CRF 模型實作法律要素標籤自動標記。

3.1 特徵值處理

由於定義的標籤性質不同，我們先標記 Global 再標記 Local labels。使用特徵爲：

- 斷詞以及詞性：由中研院斷詞系統(Hotho e*t al*., 2003)取得斷詞及詞性資訊。
- 斷詞位於判決書中的位置：以「：」、「；」以及「。」判斷句子，並進一步計算斷詞所在的句子在於整篇判決書中的位置。在實驗中，我們分別將整份判決書分爲三等份、四等份、五等份以及十等份。以三等份爲例，每個斷詞的位置資訊即代表此斷詞位於一篇判決書的前 1/3 或中 1/3 或者後 1/3。
- Cosine-normalized tfidf：對於每種標籤 k，計算動詞及名詞之特徵如下：

  $tf_{i,j}$ = 在文件 j 中，動詞或名詞 i 被標爲標籤 k 的次數

  df_i = 多少篇判決書中有動詞或名詞 i 被標爲標籤 k

$$idf_i = log\left(\frac{|D|}{df_i}\right)$$

$$tfidf_{i,j} = tf_{i,j} * idf_i$$

$$CosNorm\left(tfidf_{i,j}\right) = \frac{tfidf_{i,j}}{\sqrt{\sum_{k \in all\ labels} tfidf_{i,j}^2}}$$

標記 Local labels 時，除了以上三種特徵外，亦加入以我們系統所預測的 Global labels 值作爲特徵。

3.2 實驗結果

我們使用 CRF++(Wibowo & Williams, 2002)實作自動標記，取五疊交叉驗證(5-fold cross validation)以 F1 score 做評估，實驗結果如表 2 及表 3 所示。由實驗結果我們發現，由於 global label 通常標記一大段文字，所以判決書位置分得細標記結果反而不好；相對地當判決書位置分得越細 local label 標記結果越好。

表 2. 自動標記 Global label 結果

特徵	恐嚇取財	強盜
斷詞、詞性	0.577	0.515
斷詞、詞性、位置 1/3	0.626	0.557
斷詞、詞性、位置 1/4	0.624	0.555
斷詞、詞性、位置 1/5	0.619	0.546
斷詞、詞性、位置 1/10	0.595	0.542
斷詞、詞性、動詞名詞 Cosine-normalized TF-IDF	0.741	0.782
斷詞、詞性、位置 1/3、動詞名詞 Cosine-normalized TF-IDF	**0. 818**	**0.834**

表 3. 自動標記 Local label 結果

特徵	恐嚇取財	強盜
斷詞、詞性	0.350	0.307
斷詞、詞性、位置 1/3	0.366	0.323
斷詞、詞性、位置 1/4	0.375	0.328
斷詞、詞性、位置 1/5	0.396	0.353
斷詞、詞性、位置 1/10	0.403	0.361
斷詞、詞性、預測的 Global label 資訊	0.352	0.310
斷詞、詞性、動詞名詞 Cosine-normalized TF-IDF	0.540	0.562
斷詞、詞性、位置 1/10、動詞名詞 Cosine-normalized TF-IDF、預測的 Global label 資訊	**0.604**	**0.615**

4. 以人工標籤進行案件分類

在強盜罪與恐嚇取財罪的分類上，「行爲的脅迫程度是否足以令被害人不能抗拒」是區分兩種案件的關鍵。因此，我們利用法律要素標籤來改善分類效果，並探討影響分類結果的法律要素標籤資訊。另外，由於判決是爲法官對於某犯罪的判決紀錄，因此判決書

中的某些部份（例如「案由」）可能會提及案件所屬犯罪類型以及所判刑罰。故此，我們只取用「法官認定事實」部份作爲輸入，以求輸入資料不涉及告知犯罪類型以及判處的量刑，避免看到答案。我們首先用直覺的規則來區分強盜案件以及恐嚇取財案件：當案件事實中出現「不能抗拒」則分類此案件爲強盜案件；當案件事實中出現「心生畏懼」且無出現「不能抗拒」則分類此案件爲恐嚇取財案件，同時也考慮關鍵字前是否出現「未致」、「未達」、「不造成」等否定字。以此規則判斷分類的結果 F1 score 爲 0.852；我們以此結果做爲基準線，希望加上標籤資訊後可以有所改進。

4.1 使用特徵與實驗結果

我們使用法律要素標籤資訊以進行案件分類，使用特徵包括：

- Fact_tfidf：對於被標記爲「法官認定事實」的斷詞，計算 TF-IDF 值。
- Local label 標記次數：各 local label 被標記的次數；原始標記資料中，若有兩個斷詞標同一種 label，只以標一次計算。
- Local label 標記順序：各 local label 的標記順序值。

我們使用四疊交叉驗證(4-fold cross validation)以 F1 score 做評估，並比較兩個分類器 Liblinear 與 Logistic model tree (LMT) 的分類結果如表 4 及表 5。結果顯示使用 Fact_tfidf，local label 標記次數，和 Local label 標記順序爲特徵，並採 LMT 做分類之結果最佳，可達 F1 score 爲 0.943。另外，由 Fact_tfidf 中篩選出分類重要關鍵字如表 6。若濾掉這些關鍵字後，同樣使用 Fact_tfidf，Local label 標記次數，和 Local label 標記順序爲特徵，並採 LMT 做分類的 F1 score 降爲 0.721。因此，我們相信這些關鍵字對於案件分類是相當重要的。

表 4. Liblinear 分類結果

特徵	特徵數	F1 score
Fact_tfidf (Ftfidf)	10034	**0.834**
Local label frequency (LF)	16	0.615
Local label order (LO)	256	0.695
Ftfidf + LF	10050	0.702
Ftfidf + LO	10290	0.769
LF + LO	272	0.642
Ftfidf + LF + LO	10306	0.698

表 5. LMT 分類結果

特徵	特徵數	F1 score
Fact_tfidf (Ftfidf)	10034	0.929
Local label frequency (LF)	16	0.654
Local label order (LO)	256	0.693
Ftfidf + LF	10050	0.936
Ftfidf + LO	10290	0.921
LF + LO	272	0.678
Ftfidf + LF + LO	10306	**0.943**

表 6. 分類關鍵字

恐嚇取財	強盜
心生畏懼、恐嚇、取財、交付、晚間、拿起、意圖、取走、受有	強盜、抗拒、不能、脅迫、客觀、強暴、強取、現金、以致、得手、現場、身體、足以、強行

因此，我們選擇上述 Fact_tfidf，重要關鍵字，及其他 Local label 資訊，最後採用的特徵如表 7。爲求進一步驗證及提升結果，我們使用單一特徵、所有特徵、Leave one out (LOO)以及以 Forward selection 以找出最佳特徵組合，其實驗結果如表 8、9 及 10。從表 10 可知，最佳組合之 F1 score 爲 0.957。

表 7. 分類使用特徵

特徵	特徵描述	特徵數
Fact_tfidf (ftfidf)	標記爲「事實」區段文字之 TF-IDF 值	10034
Behavior_tfidf (btfidf)	標記爲「行爲」區段文字之 TF-IDF 值	2897
行爲關鍵字(bkey)	標記爲「行爲」的文字中是否出現表五所列之關鍵字	23
被害人反應關鍵字(vkey)	標記爲「被害人反應」的文字中是否出現表五所列之關鍵字	23
行爲人個數(actor)	1~2 人、3~5 人、6 人以上	3
行爲人特徵(afeature)	行爲人犯罪背景描述，包括無前科、有前科、有精神疾病、累犯	4
共犯個數(accomplice)	0 共犯、1~2 人、3~5 人、6 人以上	4
被害人個數(victim)	1~2 人、3~5 人、6 人以上	3
犯後態度(after)	未坦承犯行/否認犯行/未見悔意/態度不佳、坦承部分	3

	犯行、坦承犯行	
財物(property)	分爲一千以下、一千到五千、五千到一萬、一萬到十萬、十萬到百萬、錢、百萬以上、手機電子用品、本票等文件、證件、交通工具、珠寶金飾、皮包、其他	14
犯案工具(tool)	分爲電話、錄影光碟、廣義刀械、槍、道具槍、言語行爲、交通工具、棒棍、電擊棒、縱火工具、信函、掩飾衣物、綑綁工具、衣物、藥物、金屬工具、其他	17
Local label 標記次數	各 local label 在判決書中被標記的次數，原始標記資料中若兩斷詞標同一種 label 算標一次	16
Local label 標記順序	各 local label 在判決書中標記的順序	256

表 8. 使用單一及所有特徵實驗結果

特徵	**特徵數**	**F1 score**
Fact_tfidf	10034	0.929
Local label 標記次數	16	0.654
Local label 標記順序	256	0.693
行爲人個數	3	0.374
行爲人特徵	4	0.584
犯後態度	3	0.557
Behavior_tfidf	2897	0.785
行爲關鍵字	23	0.779
被害人犯影關鍵字	23	0.751
被害人個數	3	0.367
共犯個數	4	0.512
財物	14	0.603
犯案工具	17	0.670

表 9. LOO 實驗結果

特徵	特徵數	F1 score
Fact_tfidf	3263	0.850
Local label 標記次數	13281	0.921
Local label 標記順序	13041	0.936
行爲人個數	13294	0.936
行爲人特徵	13293	0.936
犯後態度	13294	0.936
Behavior_tfidf	10400	0.929
行爲關鍵字	13274	0.957
被害人反應關鍵字	13274	0.943
被害人個數	13294	0.929
共犯個數	13293	0.943
財物	13283	0.936
犯案工具	13280	0.943

表 10. 所有特徵及 Forward selection 實驗結果

特徵	特徵數	F1 score
所有特徵	13297	0.936
Fact_tfidf、Behavior_tfidf、被害人反應關鍵字、Local label 標記次數	12970	**0.957**

4.2 討論

我們希望探討每個特徵對於分類的影響，因此依據特徵值的權重來進行分析。特徵值的權重如表 11。由於前 20 項特徵絕大多數爲 TF-IDF 特徵，爲了解「非 TF-IDF 特徵」對於分類的影響，在表 12 中，我們列出分類強盜罪及恐嚇取財罪中，「非 TF-IDF 特徵」的權重。我們對於「非 TF-IDF 特徵值」做討論，並對於人工標記語料作統計佐以驗證如表 13。

表 11. 分類恐嚇取財與強盜的前 10 項特徵值權重

強盜		恐嚇取財	
權重	**特徵**	**權重**	**特徵**
16.2	ftfidf_不能	-5.22	ftfidf_恐嚇
15.41	ftfidf_抗拒	-2.84	bkey_恐嚇
12.93	ftfidf_徒手	-2.71	vkey_心生畏懼
10.7	ftfidf_旋即	-2.68	ftfidf_恐嚇
9.56	ftfidf_可信	-2.63	ftfidf_心生畏懼
9.26	ftfidf_有無	-2.58	ftfidf_否則
7.71	btfidf_抵住	-2.28	btfidf_恫稱
6.81	vkey_抗拒	-2.13	after_未坦承犯行/否認犯行/未見悔意/態度不佳
5.47	ftfidf_強盜	-2.05	ftfidf_聯絡
5.4	ftfidf_得手	-2.01	btfidf_恐嚇

表 12. 分類強盜罪及恐嚇取財罪非 TF-IDF 前 10 項特徵的權重

強盜		恐嚇取財	
權重	**特徵**	**權重**	**特徵**
6.81	vkey_抗拒	-2.84	bkey_恐嚇
4.54	bkey_抗拒	-2.71	vkey_心生畏懼
4.44	tool_槍	-2.13	after_未坦承犯行
3.42	bkey_強行	-1.57	bkey_心生畏懼
2.85	tool_刀械	-1.31	bkey_拿起
2.33	tool_道具槍	-0.68	tool_電話
1.88	bkey_強取	-0.62	bkey_撥打
1.86	bkey_取走	-0.6	vkey_交付
1.85	property_珠寶金飾	-0.59	vkey_以致
1.82	property_皮包	-0.57	tool_言語行爲

表 13. 分類特徵討論

特徵	強盜	恐嚇取財
行爲關鍵字	抗拒、強行、強取、取走、強暴、取財、得手、客觀、意圖、不能	恐嚇、心生畏懼、拿起、撥打
	使用上列關鍵字的案件中有72.3%爲強盜案件	*使用上列關鍵字的案件中有90.2%爲恐嚇取財*
被害人反應關鍵字	抗拒、不能	心生畏懼、交付、以致
	使用上列關鍵字的案件中有93.8%爲強盜案件	*使用上列關鍵字的案件中有80.4%爲恐嚇取財*
犯案工具	槍、刀械、道具槍、金屬工具(ex 鐵鋸)、換裝衣物	電話、言語行爲、信函
	使用上列工具的案件中有79.6%爲強盜案件	*使用上列工具的案件中有85.7%爲恐嚇取財*
財物	珠寶金飾、皮包、證件、其他(ex 保險箱 鑰匙)、金額一千以下、一千到五千、一萬到十萬	本票、金額十萬到百萬、百萬以上
	含有上列財物的案件中有76.6%爲強盜案件	*含有上列財物的案件中有77.3%爲恐嚇取財*
犯後態度	坦承部分犯行	未坦承犯行/態度不佳、坦承犯行
	有上列特徵的案件中有75%爲強盜案件	*有上列特徵的案件中有53.4%爲恐嚇取財*
行爲人特徵	精神疾病、無前科	
	有上列特徵的案件中有61.5%爲強盜案件	

5. 以人工標籤進行量刑預測

在案件分類後，我們希望可以進一步預測強盜案件以及恐嚇取財案件的有期徒刑刑期。訓練語料與案件分類時相同。另外，考量到「一個案件具有多個行爲人，而每個行爲人的刑期可能會不相同」的情形，我們將每個行爲人獨立成新案件；因此，訓練語料總案件數增爲 155 件。強盜罪在人工標記語料中平均判刑 66.94 個月，恐嚇取財罪在人工標記語料中平均判刑 6.96 個月。

5.1 使用特徵與實驗結果

我們分別預測強盜以及恐嚇取財案件的有期徒刑（以月爲單位），並取四疊交叉驗證(4-fold cross validation)。我們以 Pearson correlation coefficient (PCC)以及(root-mean-square error) RMSE 做評估，並使用 Additive regression 做預測。特徵除了行爲人個數外，其餘皆與案件分類時相同。

Additive regression 模型透過結合多個一維函式建立回歸表面以達到降低資料維度的目的，並且模型假設每個函式的影響皆爲累加性。假設 Y 爲預測結果，X 表獨立變數共 k 維，f爲可計算 lowess 的任意函式，則 Additive regression 模型爲下列形式

$$Y = A + \sum_{j=1}^{k} f_j\left(B_j X_j\right) + \varepsilon$$

表 14. 單一特徵預測強盜量刑結果

特徵	特徵數	PCC	RMSE
Fact_tfidf	10034	0.941	13.9824
Local label 標記次數	16	0.838	22.8940
Local label 標記順序	256	0.913	16.9355
行爲人特徵	4	0.474	36.4526
犯後態度	3	0.478	36.3674
Behavior_tfidf	2897	0.945	13.5303
行爲關鍵字	23	0.825	23.3904
被害人反應關鍵字	23	0.302	39.4695
被害人個數	3	0.295	39.5495
共犯個數	4	0.126	41.0622
財物	14	0.490	36.0889
犯案工具	17	0.607	32.9075

表 15. LOO 於預測強盜量刑結果

特徵	特徵數	PCC	RMSE
Fact_tfidf	3270	0.934	14.745
Local label 標記次數	13288	0.957	11.960
Local label 標記順序	13048	0.947	13.708
行爲人特徵	13300	0.947	13.708
犯後態度	13301	0.947	13.708
Behavior_tfidf	10407	0.940	13.708
行爲關鍵字	13281	0.949	13.615
被害人反應關鍵字	13281	0.947	13.708
被害人個數	13301	0.947	13.708
共犯個數	13300	0.947	13.708
財物	13290	0.947	13.708
犯案工具	13287	0.947	13.708

表 16. 所有特徵及 Forward selection 於預測強盜罪量刑結果

選擇特徵	特徵數	PCC	RMSE
所有特徵	13304	0.947	13.708
Behavior_tfidf、Fact_tfidf、犯案工具、財物	12961	**0.954**	**11.476**

表 17. 單一特徵預測恐嚇取財罪量刑

特徵	特徵數	PCC	RMSE
Fact_tfidf	10034	0.996	1.688
Local label 標記次數	16	0.809	11.060
Local label 標記順序	256	0.994	2.105
行爲人特徵	4	0.480	16.363
犯後態度	3	0.136	18.482
Behavior_tfidf	2897	0.996	1.776

行爲人關鍵字	23	0.947	5.998
被害人反應關鍵字	23	0.797	11.258
被害人個數	3	0.225	18.178
共犯個數	4	0.151	18.442
財物	14	0.627	14.705
犯案工具	17	0.597	14.968

表 18. LOO 預測恐嚇取財罪量刑

特徵	特徵數	PCC	RMSE
Fact_tfidf	3270	0.987	1.978
Local label 標記次數	13288	0.996	1.634
Local label 標記順序	13048	0.995	1.935
行爲人特徵	13300	0.995	1.935
犯後態度	13301	0.995	1.935
Behavior_tfidf	10407	0.995	1.935
行爲人關鍵字	13281	0.995	1.935
被害人反應關鍵字	13281	0.995	1.935
被害人個數	13301	0.995	1.935
共犯個數	13300	0.995	1.935
財物	13290	0.995	1.935
犯案工具	13287	0.995	1.935

表 19. 所有特徵及 Forward selection 於預測恐嚇取財罪量刑結果

選擇特徵	特徵數	PCC	RMSE
所有特徵	13304	0.995	1.935
Fact_tfidf、Behavior_tfidf、被害人反應關鍵字	12954	**0.996**	**1.602**

我們實驗使用單一特徵、所有特徵、Leave one out (LOO)以及以 Forward selection 找出最佳特徵組合，強盜案件之實驗結果如表 14，15 及 16，最佳組合（表 16）可達到 PCC = 0.954 及 RMSE = 11.4761 之結果。相對於平均約五年半的強盜罪刑期，我們的系

統可達到一年以下之平均預測誤差。恐嚇取財案件之實驗結果如表 17，18 及 19，最佳組合（表 19）可達到 PCC = 0.996 及 RMSE = 1.6022 之結果。相對於平均約七個月的恐嚇取財罪刑期，我們的系統可達到兩個月以下之平均預測誤差。

5.2 討論

我們依據特徵值的權重，來討論特徵值在預測量刑上的影響。特徵值的權重如表 20。爲討論「非 TF-IDF 特徵」對於量刑的影響，表 21 列出強盜及恐嚇取財案件中的「非 TF-IDF 特徵」權重。我們針對「非 TF-IDF 特徵」做討論（如表 22），並對於人工標記語料作統計以協助驗證。

表 20. 預測恐嚇取財與強盜量刑的前 10 項特徵值權重

強盜		恐嚇取財	
權重	**特徵**	**權重**	**特徵**
30.7969	bkey_強暴	56.62879	ftfidf_易科
27.9362	ftfidf_被害人	53.65348	ftfidf_此外
15.52409	ftfidf_保管	11.02504	accomplice_3~5
13.78044	ftfidf_當時	8.084648	ftfidf_犯罪
13.12859	ftfidf_便利商店	4.883647	ftfidf_起訴書
12.0484	ftfidf_喝令	3.959793	ftfidf_正當
10.61914	ftfidf_丙○	3.877146	ftfidf_開立
7.996218	ftfidf_不足採信	2.892254	ftfidf_結果
7.991696	ftfidf_收銀機	2.532131	ftfidf_當場
7.470467	ftfidf_得知	2.511225	ftfidf_得逞

表 21. 預測強盜罪及恐嚇取財罪量刑非 TF-IDF 前 10 項特徵的權重

強盜		恐嚇取財	
權重	**特徵**	**權重**	**特徵**
30.7969	bkey_強暴	11.02504	accomplice_3~5
0.555235	bkey_客觀	0.719344	victim_1~2
0.504492	bkey_拿起	6.15E-01	vkey_抗拒
0.503787	property_證件	5.80E-01	vkey_取財

0.439038	bkey_心生畏懼	5.48E-01	bkey_客觀
0.406706	tool_棒棍	5.32E-01	bkey_強行
0.38426	property_手機電子用品	0.509546	tool_槍
0.351021	property_五千到一萬	0.328142	property_一千以下
0.325931	tool_槍	0.110464	property_其他
0.322338	property_皮包	5.92E-02	bkey_恐嚇

表 22. 量刑預測特徵討論

特徵	強盜	恐嚇取財
犯後態度	*若未坦承犯行平均判刑 100.5 個月*	*若未坦承犯行平均判刑 7.26 個月*
行爲人特徵	*有前科平均判刑 97.1 個月* *有精神疾病平均判刑 53.3 個月*	*有前科平均判刑 10.35 個月* *累犯平均判刑 7.45 個月*
行爲關鍵字	*包含強暴、客觀、拿起、心生畏懼、交付、抗拒、不能、取走、脅迫、意圖等關鍵字平均判刑 81.08 個月*	*包含客觀、強行、恐嚇、身體、拿起等關鍵字平均刑期爲 11.5 個月*
被害人反應關鍵字	*包含抗拒、拿起、心生畏懼等關鍵字平均判刑 71.62 個月*	*包含抗拒、取財、心生畏懼、交付等關鍵字平均判刑 10.91 個月*
犯案工具	*使用棒棍、槍、交通工具、刀械、縱火工具等工具平均判刑 72.07 個月*	*使用槍、交通工具、刀械、信函、言語行爲、道具槍、其他等工具平均判刑 15.52 個月*
財物	*財物包含證件、手機電子用品、五千到一萬、皮包、一千以下、珠寶金飾、一千到五千、百萬以上、其他等平均判刑 72.23 個月*	*財物包含一千以下、十萬到百萬、一萬到十萬等平均判刑 9.95 個月*
共犯個數	*共犯 5 以上、1~3 人平均判刑 49 個月*	*共犯 3~5 人、1~3 人平均判刑 8.46 個月*
被害人個數	*被害人數 1~2、3~5 人平均判刑 74 個月*	*被害人數 1~2 人平均判刑 8.53 個月*

6. 以自動標記進行案件分類及量刑預測

由於人工標記法律要素標籤需大量人力與時間，因此我們以第三節中介紹之自動標記系統，取代第四節及第五節中人工標記，以分別進行案件分類及量刑預測。本研究收集之 2113 件案件中， 140 件已進行人工標記（請參考第三節），其餘 1973 件由第三節所述的自動標記系統進行標記。由於 Local label 標記結果不甚理想，因此我們僅採用「法官認定事實」以及「行爲」資訊進行案件分類及量刑預測，取四疊交叉驗證 (4-fold cross validation) 分別以 F1 score、PCC 以及 RMSE 做評估，表 23、24 分別列出自動標記在分類及量刑預測的實驗結果。

表 23. 自動標記案件分類結果

特徵	特徵數	F1 score
Fact_tfidf	10034	0.797
Behavior_tfidf	2897	0.592
行爲關鍵字	23	0.513
所有特徵	12954	**0.801**

表 24. 自動標記案件量刑預測結果

特徵	特徵數	強盜		恐嚇取財	
		PCC	RMSE	PCC	RMSE
Fact_tfidf	10034	0.073	114.251	0.019	45.812
Behavior_tfidf	2897	0.067	117.284	0.016	49.533
行爲關鍵字	23	-0.015	127.017	-0.003	51.074
所有特徵	12954	**0.085**	**112.144**	**0.022**	44.823

7. 結論

本研究針對強盜罪與恐嚇取財罪提出一自動標記法律要素標簽，並進一步分類案件與預測量刑之系統，以期能提供法官做爲判決參考。首先，我們定義強盜罪與恐嚇取財罪之法律要素標籤來闡述法律文件中的犯罪資訊，並試圖以分類器自動標記標籤。其次，透過法律要素標籤資訊，我們可以改善案件分類以及量刑預測之效果。最後，我們探討影響案件分類以及量刑預測的法律資訊，並加以統計佐證。關於本研究未來可能之延伸方向有三：(1) 我們認爲自動標記 local label 的部分，仍有極大的改善空間。(2) 考慮重疊標記以及加入法官資訊，將會有助提升案件分類及量刑預測之準確性。(3) 關於法律要素標籤資訊用於案件分類與量刑預測方面，也可佐以司法院之統計資料，以作更詳細之比較。

誌謝

特別感謝林中鶴、孫斌、潘佑達、林執中等先進爲本研究提供極大的協助。

參考文獻

Andreas, H., Staab, S. & Stumme, G. (2003). Ontologies Improve Text Document Clustering. *IEEE International Conference on Data Mining*.

Ashley, K. D. & Bruninghaus, S. (2009). Automatically Classifying Case Texts and Predicting Outcomes. *AI and Law*, 17.

Jiang, W., Wang, X.-L. & Guan, Y. (2006). Improving Sequence Tagging using Machine Learning Techniques. *International Conference on Machine Learning and Cybernetics*.

Kuo, H.-C., Tsai, T.-H. & Huang, J.-P. (2006). Building a Concept Hierarchy by Hierarchical Clustering with Join/Merge Decision. *Joint Conference on Information Sciences*.

Lame, G. (2001). A Categorization Method for French Legal Documents on the Web. *International Conference on Artificial Intelligence and Law*, 219-20.

Liu, C.-L., Chang, C.-T. & Ho, J.-J. (2004). Case Instance Generation and Refinement for Case-Based Criminal Summary Judgments in Chinese. *Journal of Information Science and Engineering*, 20, 783-800.

Schild, U. (1998). Criminal Sentencing and Intelligent Decision Support. *AI and Law*, 6, 151-202.

Wibowo, W. & Williams, H. E. (2002). Simple and Accurate Feature Selection for Hierarchical Categorisation. *ACM Symposium on Document Engineering*.

Zhang, C., Xu, X. & Zhang, C. (2008). Analysis of the Factors Affecting the Performance of CRF-based Keywords Extraction Model. *New Technology of Library and Information Service*, 24, 34-40.

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 4, December 2012, pp. 69-84

語音辨識使用統計圖等化方法

Speech Recognition Leveraging Histogram Equalization Methods

謝欣汝*+、洪志偉+、陳柏琳*

Hsin-Ju Hsieh, Jeih-weih Hung, and Berlin Chen

摘要

統計圖等化法(Histogram Equalization, HEQ)是一種概念簡單且有效的語音特徵處理技術，近年來被廣泛地研究與應用於強健性語音辨識的領域。在本論文中，我們延續統計圖等化法的研究，提出一系列使用語音特徵的空間－時間之文脈統計資訊 (Spatial-Temporal Contextual Statistics)的語音特徵強健方法；其作法是在語音之倒頻譜特徵上，利用一個簡易的差分(Differencing)和平均(Averaging)的處理方式，來得到語音特徵之文脈統計資訊後予以正規化並結合。這些新方法的作法有別於傳統之個別維度獨立正規化(Dimension-Wise)的統計圖等化法，進一步地正規化不同空間與時間之間的特徵分布資訊，因此可以降低不同聲學環境所產生的偏差，並且嘗試消除傳統之統計圖等化法無法補償的問題，亦即隨機性雜訊(Random Noise)對語音所產生的影響。本論文所有的語音辨識實驗皆是作用於國際通用的連續語音語料庫 Aurora-2 上；實驗結果顯示，我們所提出之方法相較於許多著名的特徵強化法，皆有不錯的效果。

關鍵詞：語音辨識，雜訊強健性，統計圖等化法，特徵文脈的統計

* 國立臺灣師範大學資訊工程學系 Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: hsinju@ntnu.edu.tw; berlin@ntnu.edu.tw
+ 國立暨南國際大學電機工程學系 Department of Electrical Engineering, National Chi Nan University
E-mail: jwhung@ncnu.edu.tw

Abstract

Histogram equalization (HEQ) of speech features has received considerable attention in the field of robust speech recognition due to its simplicity and excellent performance. This paper is a continuation of this general line of research, presenting a novel HEQ-based feature normalization framework which takes advantage of joint equalization of spatial-temporal contextual statistics of speech features. In doing so, we explore the use of simple differencing and averaging operations to capture the contextual statistics of feature vector components for speech feature normalization. All experiments are conducted on the Aurora-2 database and task. Experimental results show that for clean-condition training, the methods instantiated from this framework achieve considerable word error rate reductions over the baseline system, which are indeed quite comparable to other conventional methods.

Keywords: Speech Recognition, Noise Robustness, Histogram Equalization, Feature Contextual Statistics.

1. 研究動機

『科技始終來自於人性』，這是一家手機廠商的廣告用語；隨著科技不斷的進步，電腦功能不斷地提升、相關資訊設備也日漸普及並且深入到你我的日常生活中，不僅爲人類生活帶來許多的便利性，更是大大地提升工作效率與生活品質。現今我們可以藉由電腦或其它資訊設備來完成大部分的工作，如此一來便使得人類與電腦間有著密不可分的關係。但目前人類與電腦的溝通方式，仍須仰賴鍵盤、滑鼠等工具，因此對於某些特定族群的使用者而言，這種不友善的操作介面無疑是一個障礙。我們相信以最自然且簡便的方式來操作這些科技產品，能將科技帶給人們的效益提升到最高。

由於語音是人們最自然且最普遍使用的溝通媒介，因此在不久的將來，語音必然會扮演著人類與智慧型電子設備間，最重要的互動媒介，而自動語音辨識(Automatic Speech Recognition, ASR)技術將會是一個關鍵的角色。然而在現實生活中，已有許多和自動語音辨識技術相關的應用，其中最廣爲人知的應用爲航空公司的語音訂位系統及銀行帳戶的語音查詢系統等；而這一類的系統能成功運作的原因主要是因爲限制系統辨識的詞彙個數。此外還有許多的自動語音辨識相關的應用，如語音轉譯文字軟體、互動式聲音問答系統和語音文件檢索等；然而要實現這類技術，將會面臨許多困難與障礙。

對於一套自動語音辨識系統而言，在語音訊號不受雜訊干擾的理想實驗室環境下，一般皆可獲得良好的辨識結果，但若要應用至日常生活的環境中，常會受到環境中諸多雜訊的干擾，例如:具有加成性的背景雜訊(Background Noise)或是錄音設備本身所產生的摺積性的通道效應(Channel Effect)等，皆會造成系統之訓練環境與測試環境之間存在不匹配(Mismatch)的情況，而嚴重地影響系統的辨識效能。因此，在自動語音辨識技術的

發展上，雜訊強健性(Noise Robustness)一直是一門重要的研究議題。並且，如何能以更有效的方式來處理雜訊所造成的影響，將是一個既複雜又頗具挑戰性的任務。

如前所述，對於語音訊號而言，環境中雜訊的干擾大致可分爲兩種類型: (1)加成性雜訊(Additive Noise)和(2)摺積性(Convolutional Noise)雜訊。其中加成性雜訊爲錄製語音時，原始語音與背景雜訊呈線性加成的關係一同被收錄進去，例如汽車乎嘯而過或周遭人們聊天所產生的噪音等；另一方面，摺積性雜訊則是指語音訊號經由不同傳輸通道所造成的通道效應，例如麥克風通道效應、電話線路或手持式電話所產生的通道效應等。圖 1 爲乾淨語音訊號受加成性雜訊與摺積性雜訊干擾的示意圖。

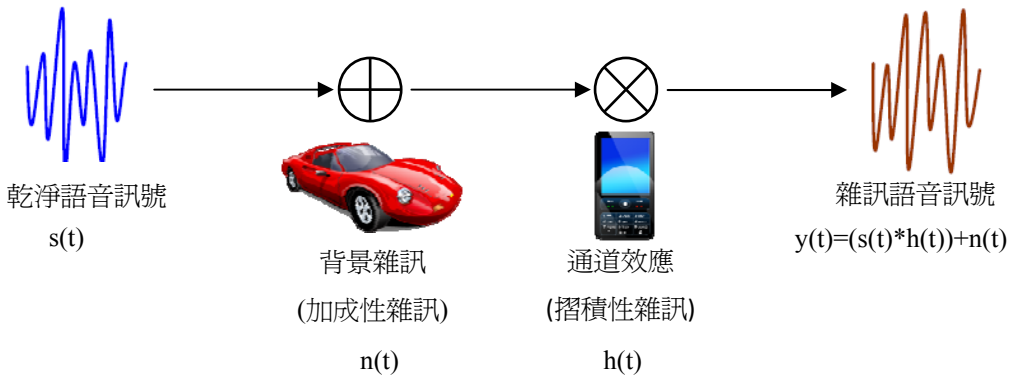


圖 1. 雜訊對語音訊號干擾示意圖

在強健性語音辨識的研究領域裡，過去已有許多學者已成功的發展出許多相關的演算法，其主要目的爲降低雜訊對語音訊號的影響，進而使得辨識結果能夠有效的提升。依據所發展出方法的特性，約略可分爲以下三大研究方向(Gong, 1995):

(1) 語音訊號增益法(Speech Enhancement):

考量人耳聽覺的特性，以增加語音訊號在感知上的品質。其主要的目的爲將語音訊號從受雜訊干擾之空間轉換至乾淨語音空間，期望轉換後的語音訊號能與對應的乾淨語音訊號相似。但此方法不保證一定可使自動語音辨識之效能提高。原因爲大多數的語音增益方法都會導致訊號失真的情形，雖然人耳對於些許訊號的失真有很好的容忍力，但是這些干擾對自動語音辨識器而言則相當敏感。常見的技術有頻譜消去法(Spectral Subtraction, SS) (Boll, 1979)、端點偵測(Voice Activity Detection, VAD) (ITU, 1996)等。

(2) 強健性語音特徵(Robust Speech Feature):

主要作法是希望從語音訊號中擷取較不易受到雜訊干擾而失真的強健性語音特徵參數，進而降低訓練語料和測試語料間存在的不匹配情況，因此可以有效的提升自動語音辨識的效能。其著名的方法有倒頻譜平均值消去法(Cepstrum Mean Subtraction, CMS) (Furui, 1981)、倒頻譜平均值與變異數正規化法(Cepstrum Mean

and Variance Normalization, CMVN) (Viikki & Laurila, 1998)與倒頻譜平均與變異數正規化法結合自動回歸動態平均濾波器法(Cepstral Mean and Variance Normalization plus Auto-Regressive-Moving Averaging Filtering, MVA) (Chen & Bilmes, 2006)等。

(3) 聲學模型調適法(Acoustic Model Adaptation):

藉由辨識器的學習，以轉換聲學模型內的分佈，進而獲得與輸入的雜訊語音向量近似的分佈。常見的技術有最大事後機率法則(Maximum a Posteriori, MAP) (Gauvain & Lee, 1994)、最大相似度線性回歸法(Maximum Likelihood Linear Regression, MLLR) (Leggetter & Woodland, 1995)、平行模型結合法(Parallel Model Combination, PMC) (Hung *et al*., 2001)等。

本論文所提出之新方法是基於上述第二類的強健性語音特徵所發展出來的。而目前最廣泛被使用的語音特徵參數包含以人耳之聽覺特性爲考量依據而發展出的梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficients, MFCCs) (Davis & Mermelstein, 1980)、線性預估倒頻譜係數(Linear Prediction Cepstral Coefficients, LPCC) (Atal, 1974)及感知線性預估倒頻譜係數(Perceptual Linear Prediction Cepstral Coefficients, PLPCC) (Hermansky, 1991)等。然而，透過這些特徵參數擷取方法所抽取出來的特徵，往往卻極爲容易受到雜訊的干擾而有所影響，而本論文所提出之方法皆是作用於梅爾倒頻譜係數的架構上。

對於以特徵爲基礎的強健性技術而言，由於和其他兩類別的強健性方法比較起來，無論是在做法上或者對於演算法之運算複雜度上，都相對比較簡易且效果十分顯著，因此目前已成功發展出一系列相關之演算法，例如倒頻譜平均值消去法(CMS)、倒頻譜平均值與變異數正規化法(CMVN)與統計圖等化法(HEQ)。基於這三種方法，可消除雜訊所造成的線性失真方法爲倒頻譜平均值消去法和倒頻譜平均值與變異數正規化法，而統計圖等化法則能補償雜訊所造成的非線性失真。本論文將此類方法歸納爲動差正規化法，將與其他種類的特徵正規化法在第二章節中給予詳盡的介紹。

本論文延續統計圖等化法的研究，提出一套新穎的語音特徵正規化技術，其作法是在語音之倒頻譜特徵上，利用一個簡易的差分和平均的處理方式，來得到原始語音特徵之相對應的文脈統計資訊後加以正規化並結合。此新方法的作法有別於傳統之個別維度獨立正規化的統計圖等化法，而是正規化不同空間與時間之間的特徵分布資訊，因此可以更進一步的降低不同聲學環境所產生的偏差，並且嘗試消除傳統之統計圖等化法無法補償的問題，即隨機性雜訊對語音所產生的影響。本論文後續安排如下：第二章節介紹一些著名的運用於時間序列之特徵正規化法的相關研究介紹；第三章則詳細介紹本論文所提出的改良式統計圖等化法，其對應之實驗結果與討論則在第四章節中呈現；最後，第五章節爲結論與未來展望。

2. 運用於時間序列之特徵正規化法的相關研究介紹

2.1 相對頻譜法(Relative Spectral, RASTA) (Hermansky & Morgan, 1994)

觀察人類發音的特性，發現其語音訊號之調變頻譜在低於 1Hz 或高於 12Hz 的範圍是屬於非語音的訊號(Non-Speech)，因此可以使用一個帶通濾波器(Band-Pass Filter)來移除非語音的成分，針對數個音框的特徵參數進行平滑的動作。此濾波器的轉移函數(Transfer Function)如下所示:

$$H_{RASTA}(z) = \frac{0.1\sum_{\theta=1}^{2}\theta(z^{\theta} - z^{-\theta})}{1-\alpha z^{-1}} \tag{1}$$

由式(1)可知此濾波器是由一差量濾波器和一無限長度脈衝響應(Infinite Impulse Response, IIR)之低通濾波器串接而成，當 $z=\alpha$ 時則產生一極點，因此可用參數 α 來控制其頻率響應之峰值所對應的頻率，且當 α 值愈大時，峰值所對應的頻率則變得更小，所以高頻部分的響應則會被壓得更低，而在本論文的辨識實驗中設為 0.94。此外，還有一個位於 0 的零點，可以有效的去除極低頻之慢速變化通道失真效應。

2.2 動差正規化法(Moment Normalization)

如前所述之倒頻譜平均值消去法、倒頻譜平均值與變異數正規化法，通常只需很少量的運算時間即可明顯提升語音辨識的效果因此被廣泛的應用，分別正規化語音特徵參數之第一階動差與第一、二階動差，其公式如下所示:

$$\overline{X}^{d} = \frac{1}{T}\sum_{t=1}^{T} x_t^d, \quad \hat{x}_t^d = x_t^d - \overline{X}^{d}, \quad 1 \le d \le D \tag{2}$$

$$\overline{X}^{d} = \frac{1}{T}\sum_{t=1}^{T} x_t^d, \quad \sigma^{d} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t^d - \overline{X}^{d})^2}, \quad \hat{x}_t^d = \frac{x_t^d - \overline{X}^{d}}{\sigma^{d}}, \quad 1 \le d \le D \tag{3}$$

其中，x_t^d 表示第 d 維的第 t 個音框的語音特徵參數，T 為總音框個數，$\overline{X}^{d}$ 和 σ^{d} 則分別代表第 d 維語音特徵的平均值(Mean)與變異數(Variance)，$\hat{x}_t^d$ 則為其分別正規化後所得之新的特徵。式(2)隱含著經由扣除平均值的處理方式，即可消除通道雜訊所造成的干擾，式(3)除了能消除通道雜訊所造成的影響，還可藉由正規化變異數的動作來降低不同維度間語音特徵分佈的差異程度，因此更能進一步的減少雜訊對語音特徵所造成的干擾。但因方法本身線性關係的限制，因此只能補償雜訊所造成之線性失真的情況，相反的此類方法對於雜訊所造成之非線性失真的情況，其補償效果則是非常有限的。此外另有學者提出正規化語音特徵的第三階動差或更高階的動差(Suk *et al.*, 1999)。

2.3 統計圖等化法(HEQ)

此方法原爲應用於影像處理的領域，以解決數位影像之色彩分佈不均、對比度不平衡等問題 (Acharya & Ray, 2005)。由於統計圖等化法是一種概念簡單且效果顯著的演算法，因此近年來已被廣泛的研究與應用於語音處理的領域中 (De la Torre *et al*., 2005; Hilger & Ney, 2006; Lin *et al*., 2009; Chen *et al*., 2011)。此方法不僅對語音特徵之平均值與變異數做正規化外，還可正規化更高階的動差。目的爲使訓練語料及測試語料的統計分佈特性能夠趨於一致，因此其效果一般而言是優於線性補償技術之倒頻譜平均值消去法及倒頻譜平均值與變異數正規化法。除此之外，容易與大部分之語音特徵表示法(Speech Feature Representation)或強健性方法結合且無需事先對雜訊做任何的假設則爲其附加的優點。主要的作法是將測試語料的機率分佈函數(Probability Distribution Function, PDF)對應至由訓練語料所統計出來的參考分布的機率密度函數，藉由此匹配轉換過程，以降低環境雜訊所造成之測試語料與訓練語料其統計特性不同的現象。

$$\hat{x}_t^d = F_{ref}^{-1}(F(x^d)), \quad 1 \le d \le D \tag{4}$$

式(4)爲統計圖等化法的轉換公式，其中 x^d 爲原始第 d 維的語音特徵參數， $F(x^d)$ 爲 x 之第 d 維的機率分佈且 $F_{ref}(\cdot)$ 爲此相同維度之參考的機率分佈，由此可知此方法是經過 D 次的獨立轉換，所得之 $\hat{x}_t^d$ 爲轉換後之新的語音特徵。

2.4 倒頻譜增益正規化法(Cepstral Gain Normalization, CGN)(Yoshizawa *et al*., 2001)

當乾淨語音受到雜訊影響之後，其雜訊語音特徵之平均值會與原始未受干擾的乾淨語音特徵值之平均值之間產生一個偏移量，同時兩者之間的動態範圍也會因爲受到雜訊的影響而產生不一致的情況，而使得辨識效果變差。因此使用倒頻譜增益正規化法即可消除上述之直流偏移量且正規化特徵參數的動態範圍，其公式如下所示:

$$\overline{X}^d = \frac{1}{T}\sum_{t=1}^{T} x_t^d, \quad \hat{x}_t^d = \frac{x_t^d - \overline{X}^d}{\max(x^d) - \min(x^d)}, \quad 1 \le d \le D \tag{5}$$

其中 max(·) 與 min(·) 分別是求出每一維原始倒頻譜特徵 x^d 之最大值與最小值的函數，而 $\overline{X}^d$ 爲原始每一維之倒頻譜特徵的平均值。

2.5 倒頻譜平均與變異數正規化法結合自動回歸動態平均濾波器法(MVA)

此方法的作法爲上述所提及之倒頻譜平均值與變異數正規化法再結合自動回歸動態平均濾波器(Chen *et al*., 2002)的處理，除了保有倒頻譜平均值與變異數正規化法的優點，利用一個 ARMA 低通濾波器(Low-Pass Filter)可消除因非穩定性雜訊(Non-Stationary Noise)所造成的異常尖峰(Sharp Peak)或波谷(Valley)並達到語音特徵的平滑化(Smoothing)、減緩音框間過度劇烈的快速變化。此外，這樣的結合方式還可進一步的改善調變頻譜平坦化的問題。其公式如下所示:

$$\hat{x}_t^d = \frac{1}{2M+1}\left(\sum_{m=0}^{M} x_{CMVN,t+m}^d + \sum_{m=1}^{M} x_{CMVN,t-m}^d\right), \quad 1 \le t \le T \tag{6}$$

其中 M 爲 ARMA 濾波器的階數，而在本論文的辨識實驗中設爲 2， x_{CMVN} 爲經過倒頻譜平均值與變異數正規化法處理後之特徵。

3. 統計圖等化法使用語音特徵的空間－時間之文脈統計資訊(ST-HEQ)

上述所提及之運用於時間序列上的強健性演算法，雖然已可達到不錯的辨識結果，但由於其並未考慮到對一語音訊號而言，不同頻率成分所造成的影響。如我們所知，對於語音辨識而言，不同頻率成分所占的重要性不盡相同且大部分的語音辨識資訊主要集中於 1Hz 至 16Hz 之間(Kanedera *et al.*, 1997)。因此我們認爲，若能進一步的對一語音訊號之不同頻率成分加以分析、處理，對語音辨識而言將會帶來更大的效益。

在語音訊號處理上，利用離散餘弦轉換(Discrete Cosine Transform, DCT)可得到接近不相關的語音特徵。例如：著名的梅爾倒頻譜特徵就是經由離散餘弦轉換後所得到的。在語音辨識的應用上，爲了要減少運算的複雜度，通常會假設特徵參數彼此間是互相無關(Uncorrelated) 的，但對數頻譜(Logarithmic Spectrum)並不符合此要求，所以將對數頻譜經由離散餘弦轉換後所得的倒頻譜特徵，彼此間的相關性大幅降低，進而較吻合特徵彼此無關的要求。因此大多數傳統的統計圖等化法，當其作用於倒頻譜域(Cepstrm Domain)時，皆是沿著語音特徵之時間序列，進行個別維度之語音特徵獨立正規化的動作，但當不同維度的語音特徵向量不是完全不相關的情形下，此一假設則是無效的。此外從短時間之語音訊號分析的觀點出發，其語音訊號的特性是隨時間緩慢變化的，因此我們假設鄰近串接的語音特徵向量能提供額外有助於正規化的資訊。另一方面，由於傳統的統計圖等化法，皆是假設雜訊對於乾淨語音的干擾是呈現單調(Monotonic)的轉換形式，但隨機性雜訊極可能會使得雜訊對於語音特徵的干擾變成非單調的轉換，此轉變將會導致無法復原的資料損失。

基於上述的觀點，本論文延續傳統統計圖等化法的研究，提出一個新穎的統計圖等化法使用語音特徵的空間－時間之文脈統計資訊；此方法不僅正規化語音特徵之整體(Overall)的統計資訊，更進一步將語音特徵之時間域(Temporal Domain)與空間域(Spatial Domain)上的局部(Local)統計資訊加以正規化。此外本方法的特點爲:突破以往傳統之統計圖等化法只考慮個別維度之統計資訊正規化的問題，並試圖藉由鄰近語音特徵向量所串接而成的文脈統計資訊來改善隨機性雜訊所造成的干擾。所提出之方法的整體概念與作法將在以下做詳細的介紹。

爲了能進一步的得到語音特徵在空間域上之不同頻率成分的文脈統計資訊。本論文將一個簡易的差分和平均的濾波器處理方式，作用於同一個音框 t 之任意兩個相鄰且不同維度的語音特徵 x_t^d 、 x_t^{d-1} ，進而將全頻帶(Full-Band)之統計圖等化法處理後的特徵一分爲二，其公式分別如下所示:

$$x_{s-diff,t}^{d}=\begin{cases}\dfrac{x_t^d-x_t^{d-1}}{2}, & 2\le d\le D\\ x_t^d, & d=1\end{cases} \tag{7}$$

$$x_{s-avg,t}^{d}=\begin{cases}\dfrac{x_t^d+x_t^{d-1}}{2}, & 2\le d\le D\\ 0, & d=1\end{cases} \tag{8}$$

其中 $x_{s-diff,t}^{d}$ 與 $x_{s-avg,t}^{d}$ 分別表示從原始語音特徵 x_t^d 之空間域上所擷取出的高頻(High-Frequency)和低頻(Low-Frequency)的語音成分。隨後將兩個分離出來的頻帶沿著其時間序列的方向，再次利用統計圖等化法來補償雜訊所造成的失真。最後，將這兩個正規化後的頻帶做線性相加的動作成新的語音特徵。相同的，將此處理方式作用於同一個維度之任意兩個相鄰的音框，亦可得到原始語音特徵 x_t^d 在時間域上之高頻 $x_{t-diff,t}^{d}$ 和低頻 $x_{t-avg,t}^{d}$ 的文脈統計資訊，且將這兩個頻帶予以正規化處理也可達到降低雜訊對語音特徵所造成的影響，其所提出之方法流程圖爲圖 2 所示。

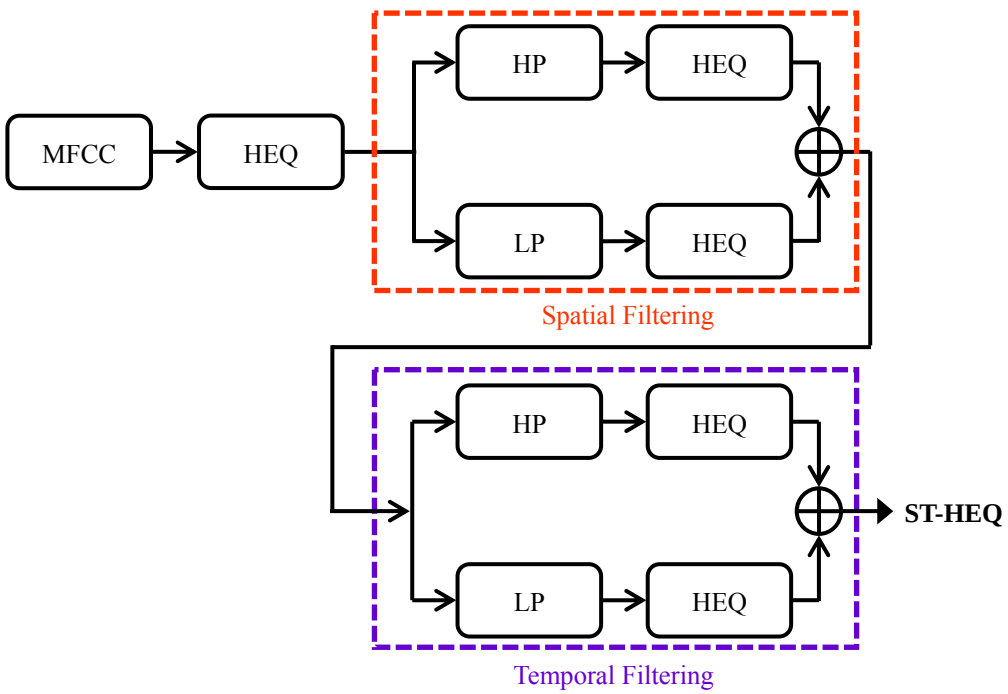


圖2. 結合式統計圖等化法使用語音特徵的空間－時間之文脈統計資訊的流程圖

如前所述，對於語音辨識而言，不同頻率成分所占的重要性不盡相同。因此本論文另外提出了一個變型的 ST-HEQ 法，稱之爲加權式統計圖等化法使用語音特徵的空間－時間之文脈統計資訊(Weighted Spatial-Temporal Contextual Statistics Histogram

Equalization, WST-HEQ)，此方法和 ST-HEQ 方法相比其主要的差別在於，ST-HEQ 於最後結合已正規化的高頻和低頻部分的語音特徵時，是直接將這兩個頻帶的特徵做線性加總的作法，而 WST-HEQ 於重建新的語音特徵時，則是利用一個變數 α 來控制語音特徵之高、低頻成分所占的比重，其重建方法如下所示:

$$\hat{x}_{all} = \hat{x}_{diff} \cdot \alpha + \hat{x}_{avg} \cdot (1-\alpha) \tag{9}$$

其中 $\hat{x}_{diff}$ 與 $\hat{x}_{avg}$ 分別代表已正規化之高、低頻語音特徵，$\hat{x}_{all}$ 爲重建之新的語音特徵且 α 爲一個小於 1 的變數。透過此作法，可進一步探討語音特徵在空間域和/或時間域上之不同頻率成分的重要性。

值得注意的是，對於語音特徵之時間域或空間域上的正規化處理方式，過去已有學者提出概念類似的語音特徵之單一域的正規化處理技術(Hung & Fan, 2009; Joshi *et al*., 2011)，而本論文所提出之結合式統計圖等化法使用語音特徵的空間－時間之文脈統計資訊的技術，於目前爲止則是相對較少被研究與探討的議題。

4. 各種強健性技術之辨識結果與討論

在本章節的一開始，會先介紹本論文在語音辨識實驗上所使用的語音語料庫，並且說明相關的實驗設定與辨識效能評估的方式，隨後會探討本論文所提出之新的語音特徵正規化方法的實驗結果，並與其他常見的特徵正規化方法作比較。此外，我們更進一步的將本論文所提出之方法與著名的 ETSI 進階前端標準(Advanced Front-End Standard, AFE) (Macho *et al*., 2002)做結合，來觀察此結合是否能更進一步地提升系統辨識的精確度。

4.1 實驗語音語料庫

爲了驗證本論文所提出的方法是否爲有效且可行的，因而在語音辨識實驗方面，我們將其作用在國際通用的連續語音語料庫 Aurora-2(ETSI, 2005)上。Aurora-2 本身爲一套含有雜訊的連續英文數字語音語料庫，其內容皆是由美國成年男女所錄製的。爲了評估各種雜訊對於語音的影響，用於測試的語料則分別夾雜了八種不同來源的加成性雜訊和兩種不同特性的通道效應。根據不同雜訊種類的干擾，進一步地將其分成三個測試集：Set A、Set B 與 Set C。其中 Set A 的語料分別含有地下鐵(Subway)、人聲(Babble)、汽車(car)和展覽會館(Exhibition)等四種加成性雜訊與 G.712 通道效應，Set B 的語料則分別含有餐廳(Restaurant)、街道(Street)、機場(Airport)和火車站(Train Station)等四種加成性雜訊與 G.712 的通道效應，此外 Set C 則分別加入了地下鐵(Subway)與街道(Street)兩種雜訊與 MIRS 通道效應。並且依語料中所含之雜訊成分的多寡，而有七種不同的訊雜比(Signal-to-Noise Ratios, SNRs)，分別爲 Clean (∞dB)、20dB、15dB、10dB、5dB、0dB 和 -5dB，其訊雜比的計算公式如下所示:

$$SNR(dB) = 10 \times \log\left(\frac{E_s}{E_n}\right) \tag{10}$$

其中 E_s 爲訊號能量而 E_n 指的是雜訊能量。Aurora-2 語音語料庫提供兩種訓練聲學模型的模式：乾淨情境訓練模式 (Clean-Condition Training) 與複合情境訓練模式 (Multi-Condition Training)，本論文統一使用乾淨語料訓練模式來進行實驗，訓練集的乾淨語句共有 8,440 句，其中並無加成性雜訊，卻包含了 G.712 的通道效應，因此在三個測試集中，訓練集只與測試集的 Set C 有通道上的不匹配。

表1. 語音特徵參數抽取設定

取樣頻率	8000 Hz
音框長度	25 ms
音框位移	10 ms
預強調濾波器	$1-(0.97)z^{-1}$
視窗類型	漢明窗
離散傅立葉點數	256點
濾波器組	共23個梅爾刻度三角濾波器
使用的特徵參數	13維MFCC(C0~C12) +13維ΔMFCC(ΔC0~ΔC12) +13維Δ^2MFCC(Δ^2C0~Δ^2C12)，共39維

4.2 實驗設定

本論文所使用的特徵參數是由 13 維(第 0 維至第 12 維)梅爾倒頻譜係數，加上其一階差量計算(Delta)及二階差量計算(Delta-Delta)，所形成總共 39 維之特徵參數，其詳細的參數設定如表 1 所示。而在訓練聲學模型的部分，則是使用劍橋大學所開發的隱藏式馬可夫模型工具(Hidden Markov Model Tool Kit, HTK)(CUED, n.d.)完成的，包含 11 個數字模型(zero, one, two, …, nine 和 oh)以及靜音模型，其中每個數字模型包含 16 個狀態且每個狀態包含 20 個高斯混合。

4.3 辨識效能的評估方式

本論文對於所有的辨識實驗而言，都是以詞正確率(Word Accuracy)來做爲評估一個演算法是否有效的依據。在此所指的詞(單字詞)則對應到每一個數字，且實驗數據皆是以百分比的方式來呈現，其計算公式爲：

$$\text{詞正確率} = \frac{\text{輸入詞總數} - (\text{取代型錯誤}+\text{插入型錯誤}+\text{刪除型錯誤})}{\text{輸入詞總數}} \tag{11}$$

值得注意的是，根據 Aurora-2 語音資料庫的設定，每一種雜訊的平均詞正確率計算方式是對於 20 dB 至 0 dB 的五種訊雜比詞正確率取平均，而排除乾淨情況與-5 dB 兩種極端的訊雜比的詞正確率，本論文後續的實驗結果之辨識率皆是依循此種呈現方式。

表2. 各種時間序列語音特徵正規化技術與改良式統計圖等化法的辨識率(%)

Method	Set A	Set B	Set C	Avg.
MFCC	54.87	48.87	63.95	54.29
ARMA	60.00	55.94	69.87	60.35
RASTA	67.44	71.90	68.45	69.43
CMS	66.81	71.79	67.64	68.97
CMVN	75.93	76.76	76.82	76.44
HEQ	80.03	82.05	80.10	80.85
CGN	80.08	81.48	80.20	80.66
MVA	80.89	82.00	81.49	81.45
S-HEQ	82.16	84.44	81.12	82.87
T-HEQ	80.42	82.53	80.73	81.33
ST-HEQ	82.52	84.90	81.81	83.33
TS-HEQ	81.85	84.41	81.11	82.72

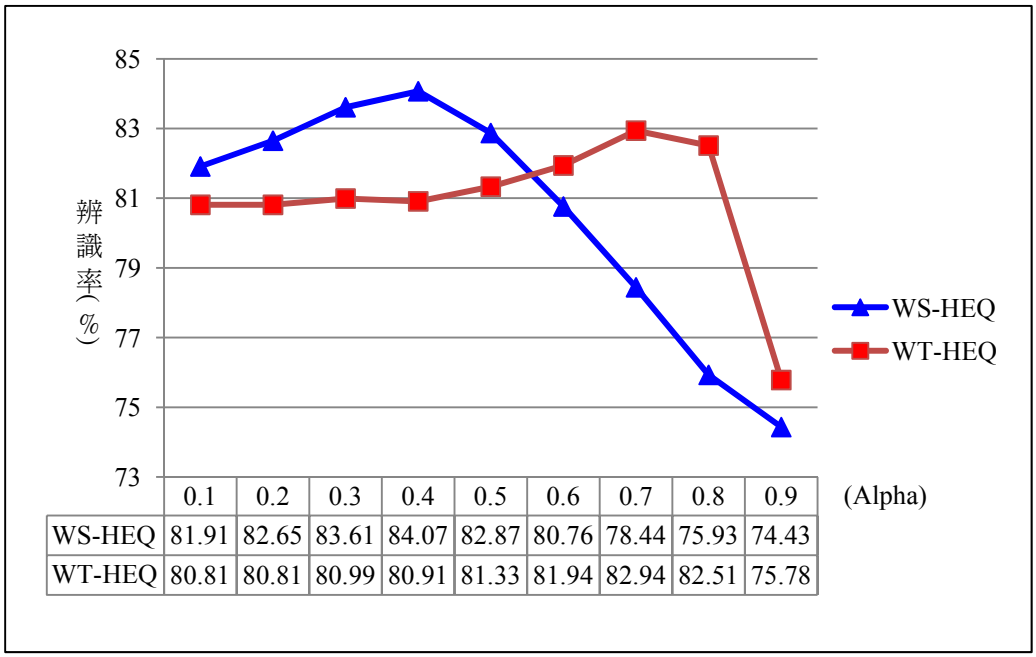
4.4 實驗結果與討論

在這一小節中，首先我們將比較本論文所提出之改良式語音特徵正規化技術(ST-HEQ)與前述所介紹之傳統的時間序列語音特徵正規化法的實驗結果，其結果如表 2 所示。隨後更進一步探討所提出之加權式統計圖等化法使用語音特徵之空間和/或時間之文脈統計資訊，對於語音辨識而言不同頻帶的重要性爲何，其結果如圖 3、4 所示。最後將 ST-HEQ 進一步結合 AFE，以便觀察這樣的結合是否有助於語音辨識率的提升，其結果如圖 5 所示。

根據表 2，我們觀察到:

1. 由於 HEQ 試圖將訓練語料和測試語料之統計分佈特性趨於一致，意謂此方法可正規化語音特徵的所有動差(All Moments)，因此可想而知的，其效果一般而言是優於 CMS、CMVN、CGN 等正規化之動差階數較少的方法。
2. 對統計圖等化法之語音特徵額外再擷取其空間和/或時間之文脈統計資訊，加以正規化後並加總，都有助於辨識效果之提升，其中 S-HEQ 的效果是優於 T-HEQ 且 ST-HEQ 優於 TS-HEQ，而最佳的正規化組合方式爲 ST-HEQ，足足可將原始辨識率從 54.29% 大幅提升至 83.33%，相對錯誤降低率約爲 64%，這顯示了此新方法對於強化語音特徵上有十分顯著的效果。
3. 對於在統計圖等化法處理後的特徵，只額外的正規化空間域上的特徵統計資訊(S-HEQ)會比額外正規化時間域上之特徵統計資訊(T-HEQ)來得好，這結果意謂由於原先已在語音特徵之時間序列域上做過一次 HEQ 處理，若額外的再做一次語音特徵

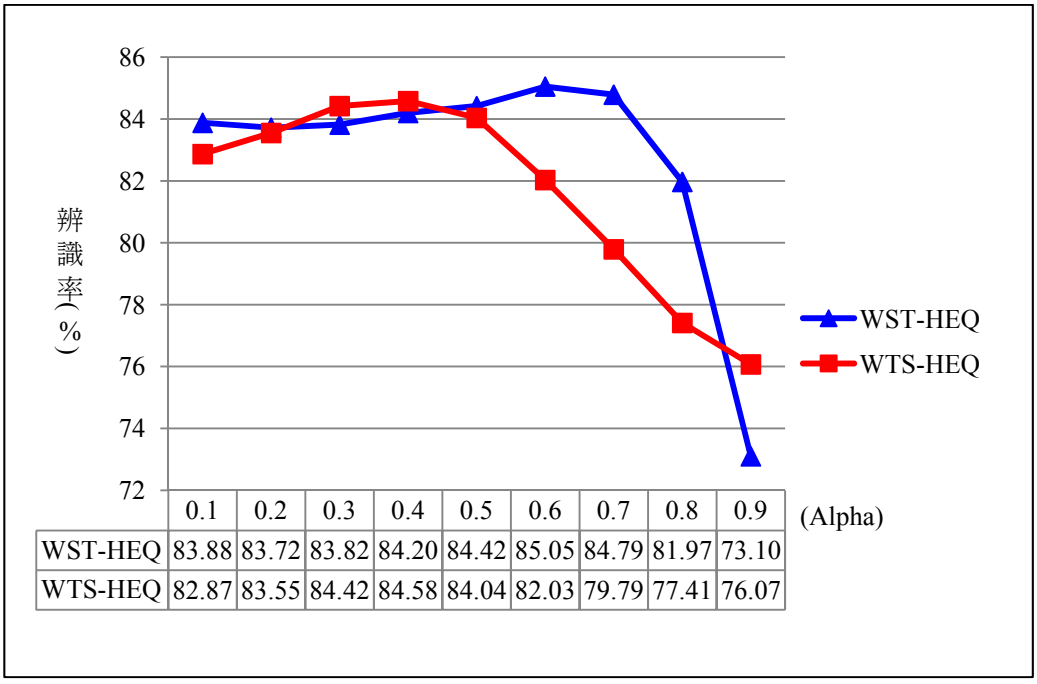
之時間域上的正規化，其效果是有限的。此時反而正規化語音特徵之空間域上的統計資訊會對語音辨識帶來更大的效益。



	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	(Alpha)
WS-HEQ	81.91	82.65	83.61	84.07	82.87	80.76	78.44	75.93	74.43	
WT-HEQ	80.81	80.81	80.99	80.91	81.33	81.94	82.94	82.51	75.78	

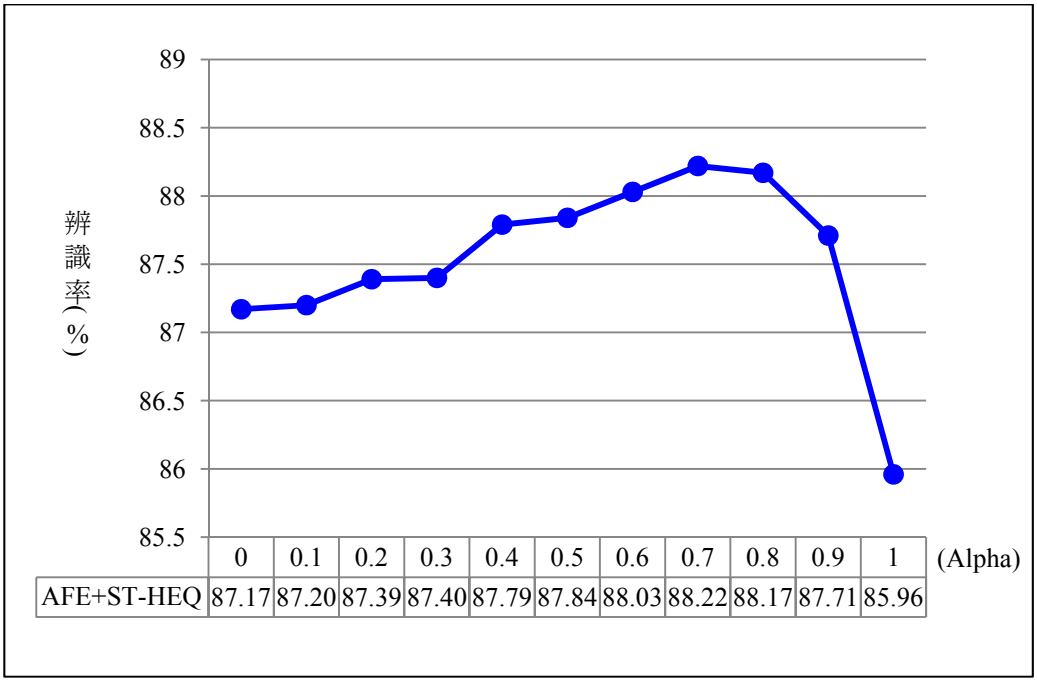
圖 3. 加權式統計圖等化法使用語音特徵的時間或空間之文脈統計資訊的辨識率(%)

圖 3 探討的是加權式統計圖等化法使用語音特徵之空間或時間之文脈統計資訊，根據圖 3 我們發現，對於 WS-HEQ 而言，當其參數 α 設爲 0.3 與 0.4 皆可使辨識率有進一步上升的效果，進步最大的幅度爲當 $\alpha = 0.4$ ，使辨識率從原始未加權的 82.87%上升至 84.07%，其絕對錯誤降低率爲 1.20%。此現象意味著進一步在統計圖等化法之語音特徵空間域上不同頻帶予以正規化，其低頻部分所包含的辨識資訊則比高頻部分所包含的還來得多。此外，對於 WT-HEQ 而言，當其參數 α 設爲 0.6 至 0.8 還可使辨識率有進一步上升的效果，進步最大的幅度爲當 $\alpha = 0.7$ 時，可使辨識率從原始未加權的 81.33%上升至 82.94%。對於此現象，我們所給予的解釋爲:對一個短時段的語音訊號而言，基於其語音能量主要集中於低頻部分的特性，在一般的寬帶雜訊環境下，理所當然的低頻區域的訊雜比會比高頻區域的訊雜比來的高，換句話說，與高頻相比，低頻區域較能提供辨識所需之資訊，相反的高頻則包含較多對辨識而言無用的雜訊成分。但在此由於已將原始語音特徵經過一次全頻帶之 HEQ 的處理即大部分的雜訊成分已被適當的補償，故此時，我們反而需要加重高頻正規化所占的比例，以藉此提高語音辨識的精確度。



	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
WST-HEQ	83.88	83.72	83.82	84.20	84.42	85.05	84.79	81.97	73.10
WTS-HEQ	82.87	83.55	84.42	84.58	84.04	82.03	79.79	77.41	76.07

圖 4. 加權式統計圖等化法使用語音特徵的時間和空間之文脈統計資訊的辨識率(%)



	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
AFE+ST-HEQ	87.17	87.20	87.39	87.40	87.79	87.84	88.03	88.22	88.17	87.71	85.96

圖 5. 進階前端(AFE)結合改良式語音特徵正規化法(ST-HEQ)之辨識率(%)

圖 4 所呈現的結果是將圖 3 所得到之最佳辨識結果的語音特徵再做一次不同域的語音特徵正規化，例如就 WST-HEQ 而言，其作法是先得到 α 設爲 0.4 的 S-HEQ 的特徵後再做一次 T-HEQ 後並嘗試找出結合後之最佳的 α 值。從圖 4 中我們發現 WST-HEQ 當其參數 α 設爲 0.4 至 0.7 時皆可提升語音辨識的精確度，其中當 $\alpha = 0.6$ 時可得到最佳的辨識結果，此外若以 WTS-HEQ 而言，當 α 設爲 0.2 至 0.5 時也可更進一步的使辨識率有所提升，且最好的參數設定爲當 $\alpha = 0.4$ 時，此外 WST-HEQ 的結果是優於 WTS-HEQ。以上實驗結果皆證明，額外的增加時間域和/或空間域上之文脈統計資訊且適度的調整已正規化之語音特徵的高、低頻成分所占的比例，將有助於語音辨識精確度的進一步提升。

AFE 是一個著名且成效非常好的一種前端之語音特徵擷取技術，從前面的實驗結果得知，本論文所提出之方法的辨識率，沒有優於 AFE，其主要的原因是 ST-HEQ 僅僅是作用於原始 MFCC 上，並沒有額外的做雜訊估測或語音增強的程序。ST-HEQ 作用在原始 MFCC 上只額外的正規化語音特徵之時間域和空間域之文脈統計資訊，如此簡易的作法即可有效的消除雜訊所造成的影響。最後我們試圖將本論文所提出之 ST-HEQ 與 AFE 做結合，來觀察所提出之新方法是否能對 AFE 帶來額外的對語音辨識有用的辨識資訊，其結果如圖五所示。由於在 AFE 的處理程序上有丟棄部分非語音音框的動作(Frames Dropping)，因此我們的結合方式是將本論文所提出之 ST-HEQ 方法直接作用於 AFE 的語音特徵上，並且和 AFE 特徵做線性加權的組合，其結果如圖 5 所示。其中當 α 設爲 1 時則爲 ST-HEQ 直接作用於 AFE 上之辨識結果，且當 α 設爲 0 時則代表 AFE 之辨識結果，從圖 5 我們發現，當 α 設爲 0.1 至 0.9 時相對於原始 AFE 的結果，都能使辨識率有進一步提升的空間，且最好的情況是發生於當 $\alpha = 0.7$ 時，其相對錯誤降低率約有 8%，此結果再次證明了我們所提出之方法的有效性。

5. 結論與未來展望

在本論文中，我們延續統計圖等化法的研究，提出了一套新穎的語音特徵正規化技術，此方法不僅能正規化語音特徵整體的統計資訊，利用一簡易的濾波器處理技術，能更進一步的對語音特徵的空間－時間之文脈統計資訊加以正規化，使得雜訊對語音訊號所造成的影響能夠大幅的降低。此外本方法的特點爲：突破以往傳統之統計圖等化法只考慮個別維度之統計資訊正規化的問題，並試圖藉由正規化鄰近語音特徵向量所串接而成的文脈統計資訊來改善傳統之統計圖等化法無法補償隨機性雜訊所造成的干擾。在國際通用的語音語料庫 Aurora-2 上，我們驗證了所提出之 ST-HEQ 法能夠大幅提升各種雜訊環境下之語音辨識的精確度。此外，我們所提出之 ST-HEQ 其辨識率都明顯高於原始 HEQ、S-HEQ、T-HEQ 和 TS-HEQ。最終更進一步的結合進階式前端標準，實驗結果顯示這樣的結合是有助於辨識效能的提升。在未來的研究中，我們嘗試將本論文所提出之演算法和其他著名的特徵正規化法做結合，來觀察辨識率是否有進一步上升的空間，並且將我們所提出之方法擴展到具有更大詞彙量的語音辨識語料庫上，以便觀察我們所提出之方法在不同複雜度之辨識系統上的效能。

致謝

本論文之研究承蒙教育部－國立台灣師範大學邁向頂尖大學計畫（101J1A0900 和 101J1A0901）與行政院國家科學委員會研究計畫(NSC 101-2221-E-003 -024 -MY3 和 NSC 99 -2221-E-003 -017 -MY3)之經費支持，謹此致謝。

參考文獻

Acharya, T. & Ray, A. K. (2005). *Image processing: principles and applications*, Wiley-Interscience.

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55, 1304-1312.

Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2), 133-120.

Chen, B., Chen, W. H., Lin, S. H. & Chu, W. Y. (2011). Robust speech recognition using spatial-temporal feature distribution characteristics. *Pattern Recognition Letters*, 32(7), 919-926.

Chen, C. P., Bilmes, J. & Kirchhoff, K. (2002). Low-resource noise-robust feature post-processing on Aurora 2.0. In *7th International Conference on Spoken Language Processing (ICSLP)*.

Chen, C. & Bilmes, J. (2006). MVA processing of speech features. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1), 257-270.

CUED. (n.d.). The hidden Markov model toolkit. Available from: http://htk.eng.cam.ac.uk.

Davis, S. B. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357-366.

De la Torre, A., Peinado, A. M., Segura, J. C., Perez-Cordoba, J. L., Benitez, M. C. & Rubio, A. J. (2005). Histogram equalization of speech representation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(3), 355-366.

ETSI. (2005). ETSI standard documentation, “Speech processing, transmission and quality aspects (STQ); distributed speech recognition; extended advanced front-end feature extraction algorithm; compression algorithms; back-end speech reconstruction algorithm”, ETSI ES 202 212 ver.1.1.2, 2005.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2), 254-272.

Gauvain, J. L. & Lee, C. H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2), 291-298.

Gong, Y. (1995). Speech recognition in noisy environments: A survey. *Speech communication*, 16(3), 261-291.

Hermansky, H. (1991). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4), 1738-1752.

Hermansky, H. & Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4), 578-589.

Hilger, F. & Ney, H. (2006). Quantile based histogram equalization for noise robust large vocabulary speech recognition. *IEEE Transactions on Speech and Audio Processing*, 14(3), 845-854.

Hung, J. W. & Fan, H. T. (2009). Subband feature statistics normalization techniques based on a discrete wavelet transform for robust speech recognition. *Signal Processing Letters, IEEE*, 16(9), 806-809.

Hung, J. W., Shen, J. L. & Lee, L. S. (2001). New approaches for domain transformation and parameter combination for improved accuracy in parallel model combination (PMC) techniques. *IEEE Transactions on Speech and Audio Processing*, 9(8), 842-855.

ITU. (1996). ITU-T Recommendation G.729-Annex B: A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70.

Joshi, V., Bilgi, R., Umesh, S., Garcia, L. & Benitez, C. (2011). Sub-band level histogram equalization for robust speech recognition. In *12th Annual Conference of the International Speech Communication Association (ICSLP)*.

Kanedera, N., Arai, T., Hermansky, H. & Pavel, M. (1997). On the importance of various modulation frequencies for speech recognition. In *European Conference on Speech Communication and Technology (Eurospeech)*.

Leggetter, C. J. & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9, 171-185.

Lin, S. H., Chen, B. & Yeh, Y. M. (2009). Exploring the use of speech features and their corresponding distribution characteristics for robust speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 17(1), 84-94.

Macho, D., Mauuary, L., Noé, B., Cheng, Y. M., Ealey, D., Jouvet, D., Kelleher, H., Pearce, D., & Saadoun, F. (2002). Evaluation of a noise-robust DSR front-end on Aurora databases. In *7th International Conference on Spoken Language Processing (ICSLP)*.

Suk, Y. H., Choi, S. H. & Lee, H. S. (1999). Cepstrum third-order normalization method for noisy speech recognition. *Electronics Letters*, 35(7), 527-528.

Viikki, O. & Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1-3), 133-147.

Yoshizawa, S., Hayasaka, N., Wada, N., & Miyanaga, Y. (2004). Cpestral gain normalization for noise robust speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1, 209-212.

The individuals listed below are reviewers of this journal during the year of 2012. The IJCLCLP Editorial Board extends its gratitude to these volunteers for their important contributions to this publication, to our association, and to the profession.

Guo-Wei Bian
Chun Chang
Tao-Hsing Chang
Ru-Yng Chang
Yu-Yun Chang
Yi-Hsiang Chao
Chao-Jan Chen
Li-mei Chen
Chia-Ping Chen
Tai-Shih Chi
Chaochang Chiu
Chih-Yi Chiu
Donghui Feng
Chih-Hsien Huang
Jeih-weih Hung
June-Jei Kuo
Yi-Chun Kuo
Wen-Hsing Lai
Bor-Shen Lin
Shu-Yen Lin
Cheng-Jye Luh
Wei-Yun Ma
Philips Kokoh Prasetyo
Wei-Ho Tsai
Gin-Der Wu
Jiun-Shiung Wu
Cheng-Zen Yang
Jui-Feng Yeh

2012 Index
International Journal of Computational Linguistics & Chinese Language Processing
Vol. 17

This index covers all technical items---papers, correspondence, reviews, etc.---that appeared in this periodical during 2012.

The Author Index contains the primary entry for each item, listed under the first author's name. The primary entry includes the coauthors' names, the title of paper or other item, and its location, specified by the publication volume, number, and inclusive pages. The Subject Index contains entries describing the item under all appropriate subject headings, plus the first author's name, the publication volume, number, and inclusive pages.

AUTHOR INDEX

Z

SUBJECT INDEX

A

B

C

D

E

F

V

W

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

1. To conduct research in computational linguistics.
2. To promote the utilization and development of computational linguistics.
3. To encourage research in and development of the field of Chinese computational linguistics both domestically and internationally.
4. To maintain contact with international groups who have similar goals and to cultivate academic exchange.

Activities：

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register：

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

payment： Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees：

regular/overseas member： NT$ 1,000 (US$50.-)
group membership： NT$20,000 (US$1,000.-)
life member：ten times the annual fee for regular/ group/ overseas members

Contact：

Address： The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

Tel.：886-2-2788-3799 ext. 1502 Fax：886-2-2788-1638

E-mail: aclclp@hp.iis.sinica.edu.tw Web Site: http://www.aclclp.org.tw

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#： _____

Name： _____ Date of Birth： _____

Country of Residence： _____ Province/State： _____

Passport No.： _____ Sex: _____

Education(highest degree obtained)： _____

Work Experience： _____

_____

Present Occupation： _____

Address： _____

_____

Email Add： _____

Tel. No： _____ Fax No： _____

Membership Category： ☐ Regular Member ☐ Life Member

Date： ____/____/____（Y-M-D）

Applicant's Signature：

Remarks： Please indicated clearly in which membership category you wish to register, according to the following scale of annual membership dues：

Regular Member ： US$ 50.- （NT$ 1,000）

Life Member ： US$500.-（NT$10,000）

Please feel free to make copies of this application for others to use.

Committee Assessment：

中華民國計算語言學學會

宗旨：

（一） 從事計算語言學之研究
（二） 推行計算語言學之應用與發展
（三） 促進國內外中文計算語言學之研究與發展
（四） 聯繫國際有關組織並推動學術交流

活動項目：

（一）定期舉辦中華民國計算語言學學術會議（Rocling）
（二）舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目
（三）收集國內外有關計算語言學知識之圖書及最新發展之資料
（四）發行有關之學術刊物，論文集及通訊
（五）研定有關計算語言學專用名稱術語及符號
（六）與國際計算語言學學術機構聯繫交流
（七）其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會
2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
   信用卡：請至本會網頁下載信用卡付款單

年費：

終身會員：	10,000.-	（US$ 500.-）
個人會員：	1,000.-	（US$ 50.-）
學生會員：	500.-	（限國內學生）
團體會員：	20,000.-	（US$ 1,000.-）

連絡處：

地址：台北市115南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw 網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

中華民國計算語言學學會
個人會員入會申請書

<table>
<tr><td>會員類別</td><td colspan="2">□終身 □個人 □學生</td><td>會員編號</td><td colspan="2">（由本會填寫）</td></tr>
<tr><td rowspan="2">姓　　名</td><td rowspan="2"></td><td rowspan="2">性別</td><td rowspan="2"></td><td>出生日期</td><td>年　月　日</td></tr>
<tr><td>身分證號碼</td><td></td></tr>
<tr><td>現　　職</td><td></td><td>學　　歷</td><td colspan="2"></td></tr>
<tr><td>通訊地址</td><td colspan="5">□□□</td></tr>
<tr><td>戶籍地址</td><td colspan="5">□□□</td></tr>
<tr><td>電　　話</td><td colspan="2"></td><td>E-Mail</td><td colspan="2"></td></tr>
<tr><td colspan="6">申請人:　　（簽章）

中　華　民　國　　年　　月　　日</td></tr>
</table>

審查結果:

1. 年費：

   終身會員：　10,000.-
   個人會員：　1,000.-
   學生會員：　500.-（限國內學生）
   團體會員：　20,000.-

2. 連絡處：

   地址：台北市南港區研究院路二段128號 中研院資訊所(轉)
   電話：(02) 2788-3799　ext.1502 傳真：(02) 2788-1638
   E-mail：aclclp@hp.iis.sinica.edu.tw　　網址: http://www.aclclp.org.tw
   連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)

PAYMENT FORM

Name :_____(Please print) Date: _____

Please debit my credit card as follows: US$_____

❑ VISA CARD ❑ MASTER CARD ❑ JCB CARD Issue Bank:_____

Card No.: _____- _____ - _____ - _____ Exp. Date:_____

3-digit code: _____ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE : _____

Tel.: _____ E-mail: _____

Add: _____

PAYMENT FOR

US$ _____ ❑ Computational Linguistics & Chinese Languages Processing (CLCLP)

Quantity Wanted: _____

US$ _____ ❑ Publications:_____

US$ _____ ❑ Text Corpora: _____

US$ _____ ❑ Speech Corpora:_____

US$ _____ ❑ Others: _____

US$ _____ ❑Life Member Fee ❑ New Member ❑Renew

US$ _____ = Total

Fax : 886-2-2788-1638 or Mail this form to :

ACLCLP

% Institute of Information Science, Academia Sinica

R502, 128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan

E-mail: aclclp@hp.iis.sinica.edu.tw

Website: http://www.aclclp.org.tw

中 華 民 國 計 算 語 言 學 學 會
信用卡付款單

姓名: _____(請以正楷書寫)　日期:： _____

卡別：❑ VISA CARD ❑ MASTER CARD ❑ JCB CARD　發卡銀行：_____

卡號:_____-_____-_____- _____　有效日期：_____

卡片後三碼：_____（卡片背面簽名欄上數字後三碼）

持卡人簽名：　_____(簽名方式請與信用卡背面相同)

通訊地址：_____

聯絡電話：_____　E-mail：_____

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$_____❑ 中文計算語言學期刊(IJCLCLP)

NT$_____❑ 中研院詞庫小組技術報告

NT$_____❑ 中文（新聞）語料庫

NT$_____❑ 平衡語料庫

NT$_____❑ 中文詞庫八萬目

NT$_____❑ 中文句結構樹資料庫

NT$_____❑ 平衡語料庫詞集及詞頻統計

NT$_____❑ 中英雙語詞網

NT$_____❑ 中英雙語知識庫

NT$_____❑ 語音資料庫_____

NT$_____❑ 會員年費　❑續會　❑新會員　❑終身會員

NT$_____❑ 其他: _____

NT$_____= 合計

填妥後請傳真至 02-27881638 或郵寄至:
115台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

		Surface	AIR (US&EURP)	AIR (ASIA)	VOLUME	AMOUNT
1.	no.92-01, no. 92-04(合訂本) ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US$ 9	US$ 19	US$15	_____	_____
2.	no.92-02 V-N 複合名詞討論篇 & 92-03 V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05 中文詞類分析	10	15	13	_____	_____
7.	no.93-06 現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02 古漢語字頻表	11	16	14	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12.	no.95-03 訊息爲本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01 古漢語詞頻表（甲）	19	31	25	_____	_____
15.	no.97-02 論語詞頻表	9	14	12	_____	_____
16.	no.98-01 詞頻詞典	18	30	26	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 20 per copy)	---	100	100	_____	_____
21.	Readings in Chinese Language Processing	25	25	21	_____	_____
				TOTAL	_____	_____

10% member discount: _____Total Due:_____

- **OVERSEAS USE ONLY**
- PAYMENT： ☐ Credit Card (Preferred)
  ☐ Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or “中華民國計算語言學學會”
- E-mail：aclclp@hp.iis.sinica.edu.tw

Name (please print): _____ Signature: _____

Fax: _____ E-mail: _____

Address：_____

中華民國計算語言學學會

相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本)　ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05　中文詞類分析	185	205	_____	_____
7.	no.93-06　現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02　古漢語字頻表	180	200	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03　訊息爲本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01　古漢語詞頻表（甲）	400	450	_____	_____
15.	no.97-02　論語詞頻表	90	100	_____	_____
16	no.98-01　詞頻詞典	395	440	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義（中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊（一年四期）　年份：_____（過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合　計	_____	_____

※ 此價格表僅限國內 (台灣地區) 使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人： 黃琪 小姐、何婉如 小姐　　E-mail:aclclp@hp.iis.sinica.edu.tw

訂購者：_____ 收據抬頭：_____

地　　址：_____

電　　話：_____ E-mail: _____

Information for Authors

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites submission of original research papers in the area of computational linguistics and speech/text processing of natural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions should report significant new research results in computational linguistics, speech and language processing or new system implementation involving significant theoretical and/or technological innovation. The submitted papers are divided into the categories of regular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full details. Short papers can focus on a smaller research issue. And survey papers should cover emerging research trends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no strict length limitation on the regular and survey papers. But it is suggested that the manuscript should not exceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced A4 pages. All contributions will be anonymously reviewed by at least two reviewers.

Copyright : It is the author's responsibility to obtain written permission from both author and publisher to reproduce material which has appeared in another publication. Copies of this permission must also be enclosed with the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to facilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright form, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP society, will be required before the manuscript can be accepted for publication. The papers published by IJCLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database services.

Style for Manuscripts: The paper should conform to the following instructions.

1. Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size of 11 points or larger.

2. Title and Author: The first page of the manuscript should consist of the title, the authors' names and institutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and e-mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such other words as the orthography of the language requires beginning with a capital letter. The author's name should appear below the title.

3. Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords is required. The abstract should not only indicate the scope of the paper but should also summarize the author's conclusions.

4. Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-hand margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).

5. Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with superscript numbers. Footnotes may appear at the end of manuscript

6. Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or written clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses.

7. References: All the citations and references should follow the APA format. The basic form for a reference looks like

Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article. *Title of Periodical, volume number*(issue number), pages.

Here shows an example.

Scruton, R. (1996). The eclipse of listening. *The New Criterion, 15*(30), 5-13.

The basic form for a citation looks like (Authora, Authorb, and Authorc, Year). Here shows an example. (Scruton, 1996).

Please visit the following websites for details.

(1) APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)

(2) APA Stytle (http://www.apastyle.org/)

No page charges are levied on authors or their institutions.

Final Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply final manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

Online Submission: http://www.aclclp.org.tw/journal/submit.php

Please visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.17 no.4 December 2012

Contents

Special Issue Articles:
Selected Papers from ROCLING XXIV

Papers

ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing