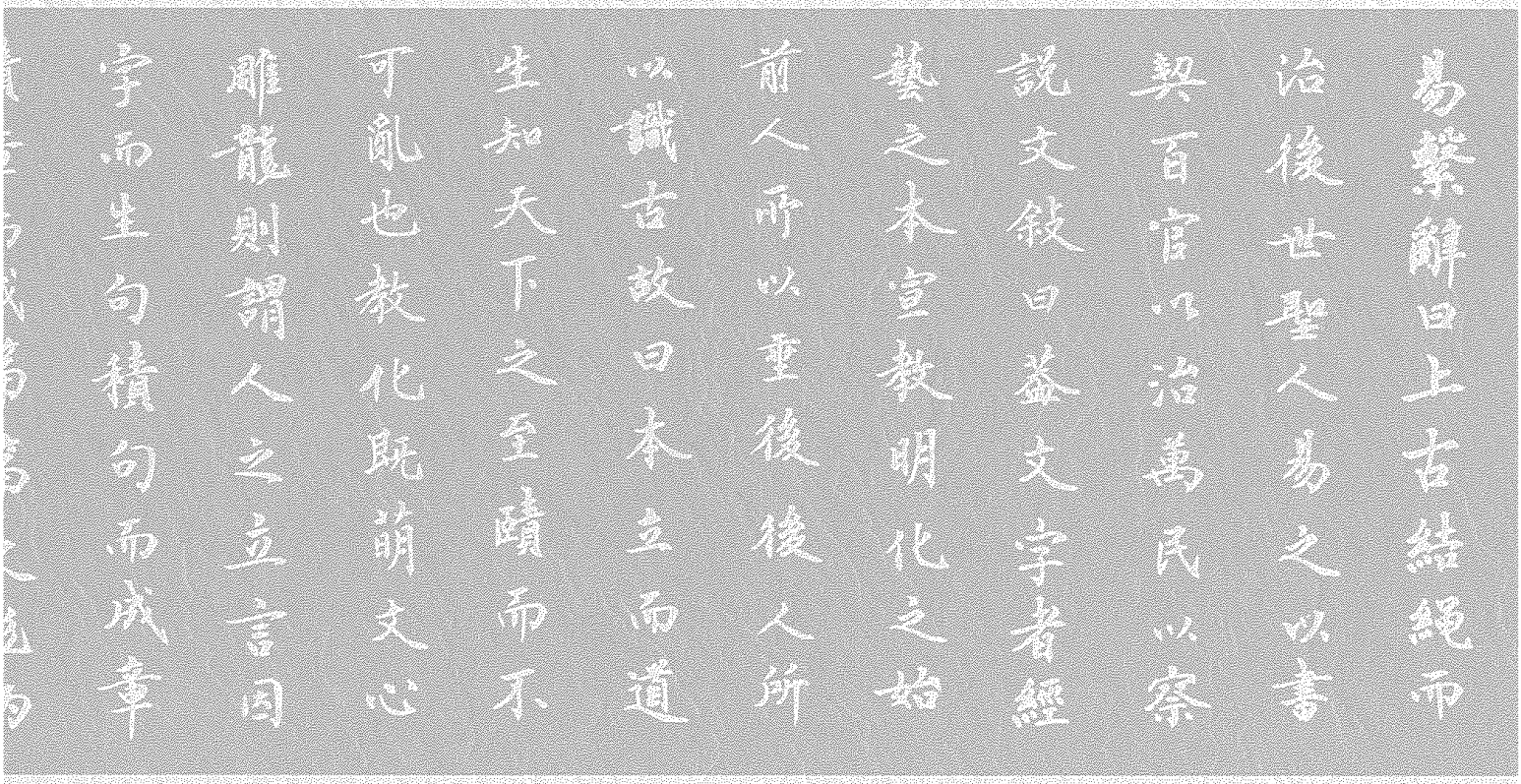
International Journal of

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.



Vol.18 No.1 March 2013 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. 18, no. 1, March 2013

Contents

Papers

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 1, March 2013, pp. 1-18

Lexical Coverage in Taiwan Mandarin Conversation

Shu-Chuan Tseng*

Abstract

Information about the lexical capacity of the speakers of a specific language is indispensible for empirical and experimental studies on the human behavior of using speech as a communicative means. Unlike the increasing number of gigantic text- or web-based corpora that have been developed in recent decades, publicly distributed spoken resources, espcially conversations, are few in number. This article studies the lexical coverage of a corpus of Taiwan Mandarin conversations recorded in three speaking scenarios. A wordlist based on this corpus has been prepared and provides information about frequency counts of words and parts of speech processed by an automatic system. Manual post-editing of the results was performed to ensure the usability and reliability of the wordlist. Syllable information was derived by automatically converting the Chinese characters to a conventional romanization scheme, followed by manual correction of conversion errors and disambiguiation of homographs. As a result, the wordlist contains 405,435 ordinary words and 57,696 instances of discourse particles, markers, fillers, and feedback words. Lexical coverage in Taiwan Mandarin conversation is revealed and is compared with a balanced corpus of texts in terms of words, syllables, and word categories.

Keywords: Taiwan Mandarin, Conversation, Frequency Counts, Lexical Coverage, Discourse Items.

1. Introduction

Exchange and communication of thoughts are mainly performed by producing and perceiving/interpreting words, whether in text or speech. In spite of philosophical debates on the concept of words, it is more or less accepted by most of the disciplines working with languages that one of the possibilities of exploring the lexical capacity of the users of a specific language is to examine the distribution of words collected in a large-scale balanced corpus. Different from the lexical entries listed in a dictionary, corpus data provide

* Institute of Linguistics, Academia Sinica, Taipei, Taiwan
E-mail: tsengsc@gate.sinica.edu.tw

information about lexical knowledge of language users that resembles their experiences and abilities in a realistic context. Of this information, word frequency counts are simple and primitive information. Nevertheless, they are directly associated with the lexical capacity of language users in a given scenario. Word frequency is one of the most essential kinds of information when implementing language-related technology tools and systems. Once a reliable word list is available, different computational models can be developed or applied to examine the role lexical knowledge plays in using a language (Baayan, 2001). For pedagogical purposes, word counts based on real corpus data will help prepare authentic learning materials for first and second language learners (Xiao *et al*., 2009; Knowles, 1990; McCarthy, 1999). For research purposes, empirical information about lexical capacity is indispensible for constructing stimuli and testing hypotheses for word- or phonology-related psycholinguistic experiments (Wepman & Lozar, 1973). In each kind of application using the word distribution information mentioned above, it is important that the sources we obtain the information from should resemble the word distribution of tokens and types as authentic language input available to the language users.

Nearly a century ago, Thorndike (1921) listed the 10,000 most widely used English words based on a 4.6-million-word corpus consisting of 41 different sources, which included children's literature, the Bible, classics, elementary school textbooks, and newspapers. The later version extended the list to 30,000 words (Thorndike & Lorge, 1944). The main purpose of these earliest wordlists was to provide word information for teaching English. Nowadays, taking advantage of the latest technology, the amount and scale of textual corpora being collected via digital resources in recent decades have become enormous. The British National Corpus (BNC) contains 100 million English words. Within the corpus data, 90% were based on written texts (Leech *et al*., 2001). The first released version of the American National Corpus (ANC) contained 11.5 million English words, 70% of which were written texts (Reppen & Ide, 2004). Both the BNC and the ANC are balanced corpora. They consist of texts collected from different producers and genres, also including transcripts of spoken language. Purely textual corpora, such as the English Gigaword and the Chinese Gigaword, distributed by the Linguistic Data Consortium (LDC), are mostly collections of newspaper articles, reflecting a specific kind of language user behavior. Nevertheless, to reflect the lexical capacity of language users in natural speech communication, we need a corpus of "naturally produced" conversations with different sociolinguistic designs of speaker relationships and different conversation types. Compared with textual corpora, however, it is considerably more difficult to obtain this kind of corpora.

Collecting and processing speech data cannot be accomplished automatically. The cost of preparing spoken corpora is high, especially when dealing with natural conversations. The types of spoken corpora vary to a large degree, ranging from reading a list of words/texts,

telling a story, executing a task, to free conversation. To take English as an example, a number of conversational corpora have been collected for educational, clinical, or experimental studies of spoken word distribution (French *et al*., 1930; Howes, 1964; Howes, 1966). They have attracted intensive attention, because they provide the most realistic materials to study how people converse to exchange thoughts and perform communication. During the last twenty years, the scale and the application of spoken corpora have been enormously extended. Svartvik and Quirk (1980) published a corpus of English conversation, later known as the London-Lund Corpus of English Conversation. A word frequency count of 190,000 words from the corpus was published four years later (Brown, 1984). Later, a part of the BNC also contained conversations, with a focus on a balanced socio-geographic sampling of speakers of English (Crowdy, 1993).

With the growing number of spoken corpora being or having been processed, the technology and the concept of how to prepare spoken corpora has also been changed accordingly due to the extensive application possibilities and the available software (Gibbon *et al*., 1997). Newly developed spoken corpora, for instance, transcribed with annotation schemes marking targeted linguistic phenomena, time-aligned with speech signals at different linguistic levels, automatically processed for word segmentation and parts of speech tagging on the transcripts, etc., have brought new horizons of how spoken corpora can be used for academic and educational purposes.

2. Taiwan Mandarin Spoken Wordlist

This paper studies the lexical coverage of a Taiwan Mandarin conversational corpus based on the derived *Taiwan Mandarin Spoken Wordlist* and compares it with the Sinica Corpus (Chen & Huang, 1996), which is currently the largest POS-tagged text corpus of Taiwan Mandarin. This section gives an introduction to how the conversational corpus has been collected and processed and how the wordlist has been prepared.

2.1 Taiwan Mandarin Conversational Corpus

The Taiwan Mandarin Conversational Corpus (the TMC Corpus, hereafter) is composed of three sub-corpora of Taiwan Mandarin conversations, which have been processed at the Institute of Linguistics, Academia Sinica (Tseng, 2004). The Mandarin Conversational Dialogue Corpus (the MCDC) is a collection of 30 free conversations between speakers who were meeting for the first time (37 females and 23 males, with ages between 16 and 45). The project was executed in 2001. One year later, 30 speakers from the MCDC speakers were recruited again to record conversations with a person they knew well for the next two corpus collection projects. As a result, 33 female and 27 male speakers whose age ranged from 14 to 63 participated in the project. The Mandarin Topic-oriented Conversation Corpus (the MTCC)

is a collection of topic-specific conversations on selected news or events that took place in the year of 2001. The Mandarin Map Task Corpus (the MMTC) is a collection of task-oriented dialogues, basically following the Map Task design (Anderson *et al*., 1991). Different from the MTCC and the MMTC, the free conversations in the MCDC were more formal, as the conversation partners were strangers. The final version of the TMC Corpus consists of 85 conversations, approximately 42 hours of speech recording. Five conversations were not included in the TMC Corpus because the participants spoke Taiwan Southern Min instead of Taiwan Mandarin to their conversation partners most of the time in their conversations. General information about the corpora is summarized in Table 1.

Table 1. Corpus Description of the TMC Corpus.

Sub-Corpus	No. of Speakers	Length per conversation	Corpus Scenario	Conversation partners
MCDC	60 (37F, 23M)	1 hour	Free conversation	Strangers
MTCC	58 (33F, 25M)	20 minutes	Topic-oriented Conversation	Friends/ relatives
MMTC	52 (28F, 24M)	7 minutes	Map task dialogue	Friends/ relatives

From the viewpoint of speaker relationship, the TMC Corpus contains conversations between strangers and conversations between people who are familiar with each other. From the viewpoint of the speaking situation, the TMC Corpus includes three different scenarios: free conversations, topic-specific conversations, and task-oriented conversations. That is, the TMC Corpus provides speech data of a variety of speaker groups communicating in different speaking styles and situations.

2.2 Corpus Transcription

The speech content of the 85 conversations was orthographically transcribed and carefully cross-checked. Words were transcribed in traditional Chinese characters. Pauses and paralinguistic sounds, such as inhalation, coughing, and laughter, were indicated in the transcripts. Items that are often used in spoken discourse, such as discourse particles, discourse markers, fillers, and feedback words, were transcribed with capital letters for two reasons. On the one hand, we wanted to distinguish these items from ordinary words due to their pragmatic function in conversation. On the other hand, it is not always possible to find the correct, or widely accepted, characters to transcribe these groups of items. For example, well-conventionalized characters are available in the writing notion for most of the discourse particles (Chao 1965) originating from Mandarin Chinese, such as: **A** 啊, **AI YA** 哎呀, **AI YOU** 唉呦, **BA** 吧, **E/EP** 呃, **EN** 嗯, **HAI** 嗨, **HE** 呵, **HEI** 嘿, **HWA** 嘩, **LA** 啦, **LIE/LEI** 咧, **LO** 囉, **MA** 嘛,

NOU/NO喏, **O**喔/噢/哦, **OU**噢, **WA**哇, **WA SAI**哇塞, **YE**耶, **YI** 咦, and **YOU**呦. Nevertheless, some of the very common particles in contemporary Taiwan Mandarin conversation, such as **EIN**, **HAN**, **HEIN**, **HO**, **HYO**, and **HAIN,** originate from Taiwanese Southern Min - a major dialect spoken in Taiwan. For these particles, no widely acceptable characters are available to transcribe them. Capital letters signifying the way of pronunciation were used to transcribe discourse particles of this kind. Different from discourse particles, discourse markers noted in our transcribing system are originally lexical items, i.e. regular words with a matching character in the writing system. When their original semantic meaning is lost and their use becomes essentially pragmatic in conversation, however, they are regarded as a kind of discourse markers. Their function is similar to that of the discourse markers that are generally defined, e.g. *well*, *but*, and *ok* (Schiffrin, 1988), marking emerging structure of conversation. In principle, they are used for a speaker to keep the floor or to stall more time to think of what to say next. Among the discourse markers annotated in the TMC Corpus, **NA** is the most frequently used marker. Originally, 那 (**NA)** was a demonstrative determiner, meaning "that". As a discourse marker, however, it sometimes appears before a proper noun, which is grammatically incorrect in the case of a determiner. This example illustrates the difference between 那 (**NA)** as a determiner and as a discourse marker. As a result, we noted discourse markers of this specific group, including **NA**, **NE**, **NA GE**, **NE GE**, **NEI GE**, **SHEN ME**, and **ZHE GE**.

The third type, fillers and feedback words, themselves do not involve any concrete semantic meaning. Fillers function as discourse markers in a similar way, indicating hesitations in speech flow (Shriberg, 1994). Feedback words are used as a response signal to the conversation partner. Different foci on spoken discourse may lead to diversified terminologies and systems of lexical items, for instance, Chao (1965) may regard some of the fillers and feedback words as interjections, carrying specific intonation contours. Nevertheless, in the TMC Corpus, our preliminary goal was to develop a coherent transcription convention for conversation. Basically, we transcribed them according to their syllable structure, because the surface forms of fillers and feedback words are systematically similar. Prosodic realization may add affined pragmatic interpretations to fillers and feedback words. Nevertheless, in the transcription system, we do not make further distinctions. There are four different sub-groups of fillers and feedback words: Zero onset + Schwa + dental nasal coda (**UHN**, **UHNN**, **UHNHN**), zero onset + Schwa + bilabial nasal coda (**UHM**, **UHMM**, **UHMHM**), dental nasal onset + Schwa + dental nasal coda (**NHN**, **NHNN**, **NHNHN**), and bilabial nasal onset + Schwa + bilabial nasal coda (**MHM**, **MHMM**, **MHMHM**, **MHMHMHM**, **MHMHMHMHM**). When they are produced with more than one syllable, each syllable is presented by a repeated **H**. A repeated nasal coda indicates a prolongation of the coda.

Foreign words, such as English or Japanese, are either written in their original writing convention or the equivalent romanization. Speech stretches containing pronunciation variants and code switching are transcribed in the way that the meaning of the speech content is written in Taiwan Mandarin writing convention.

2.3 Time-aligned Transcripts in PRAAT

The orthographic transcription of the corpus is presented in PRAAT with two tiers (Boersma & Weenink, 2012). The first tier gives information about the speaker identity and the sequence number of the speaker's turn in a coded way, and the transcription of the speech content is presented on the second tier. The boundaries of all speaker turns are time-aligned with the speech signal. **Figure 1** is an extract from the MCDC sub-corpus.

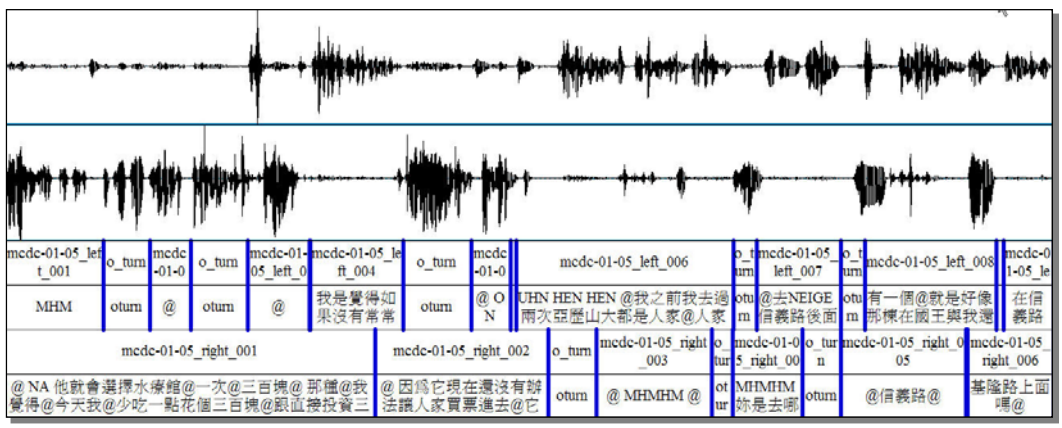


Figure 1. Time-aligned transcription.

2.4 Word Segmentation and POS Tagging

Word boundaries in the Chinese texts are not marked by blanks. In order to prepare the wordlist of the TMC Corpus, we applied the CKIP word segmentation and POS tagging system to automatically process the transcripts (Chen & Huang, 1996). The POS tagset developed by the CKIP team is listed in **Table 2** (CKIP, 1998). Slightly modifying the tagset, we added nominal expressions and idioms to the category S, because they act as independent sentences in conversation from both syntactic and pragmatic points of view and they should not be regarded as any one of the other POS categories. With regard to the input format of the system, the original design of the CKIP system was sentences. For processing the TMC Corpus, the content of each speaker turn was used as individual input to run the CKIP system. As the majority of the corpus data are long speaker turns of more than one sentence, there may arise difficulties in word segmentation and POS tagging. In this regard, manual post-editing would be necessary.

Table 2. The CKIP POS Tagset.

Word category	CKIP POS Tagging system
Adjectives	Non-predicative adjective (A)
Adverbs	Adverb (D), quantitative adverb (Da), pre-verbal adverb of degree (Dfa), post-verbal adverb of degree (Dfb), sentential adverb (Dk), aspectual adverb (Di)
Conjunctions	Coordinate conjunction (Caa), correlative conjunction (Cbb), conjunction:*deng3deng3* (Cab), conjunction:*de5hua4* (Cba)
Determinatives	Demonstrative determinatives (Nep), quantitative determinatives (Neqa), specific determinatives (Nes), numeral determinatives (Neu), post-quantitative determinatives (Neqb)
Foreign words	Foreign words (FW)
Interjections	Interjection (I)
Nouns	Measure (Nf), common noun (Na), proper noun (Nb), place noun (Nc), localizer (Ncd), time noun (Nd), postposition (Ng), nominalization (Nv)
Particles	Particle (T)
Prepositions	Preposition (P)
Pronouns	Pronoun (Nh)
Sentence	Nominal expression, idioms (S)
Verbs	Active intransitive verb (VA), active pseudo-transitive verb (VB), stative intransitive verb (VH), stative pseudo-transitive verb (VI), active causative verb (VAC), active transitive verb (VC), active verb with a locative object (VCL), ditransitive verb (VD), active verb with a sentential object (VE), active verb with a verbal object (VF), classificatory verb (VG), stative causative verb (VHC), stative transitive verb (VJ), stative verb with a sentential object (VK), stative verb with a verbal object (VL), you3 (V_2)
DE	Structural particles: *de5, zhi1, de2, di4*
SHI	Copula: *shi4*

2.5 Manual Post-editing of Word Segmentation and Homograph Errors

The CKIP word segmentation system was originally trained on written texts. Therefore, incomplete, ungrammatical sentences and peculiar constructions in conversation, which normally do not occur in written texts, could result in errors of the automatic word segmentation and POS tagging system. Segmentation errors, including errors of proper nouns, idioms, constructions with numbers, and directional complements, were manually corrected. In the process of word segmentation and POS tagging, we also need to cope with the occurrences of disfluencies in conversation (Shriberg, 1994). According to the content and the

prosodic realization, a disfluent repetition of words was manually separated (e.g. da uhn da de jiqi, *big uhn big machine*), whereas a grammatical reduplicative phrase was transcribed as one unit (e.g. dadade chengzan ta, *a big compliment to him*).

To obtain information about syllables, all Chinese characters transcribed were automatically converted into Hanyu Pinyin, a romanization convention for Chinese used worldwide. In the system of Hanyu Pinyin, tone information is included with each syllable, which is indicated by 1, 2, 3, 4, and 5, representing Tone 1, Tone 2, Tone 3, Tone 4, and the neutral Tone. Furthermore, because of the large number of homographs in Chinese, post-editing was performed to manually correct errors resulting from the automatic conversion. Ambiguous homographs, which occur very frequently in spoken language, were specified based on the neighboring context. For instance, the word "*one*" (一, yi1) is pronounced with Tone 1 in isolation, but with Tone 2, when followed by Tone 4 and the neutral tone. When followed by Tone 1, Tone 2, and Tone 3, 一 is pronounced with Tone 4. The final version of the automatically segmented and POS tagged words, as well as the manually checked syllables, was used to prepare the wordlist. As a result, the *Taiwan Mandarin Spoken Wordlist*[1] contains 405,435 regular word tokens, equivalent to 16,683 word types and 607,008 syllable tokens. There are 57,696 tokens of discourse particles, discourse markers, fillers, and feedback words.

3. Lexical Coverage in Conversational and Text Corpus

Given a body of language data, no matter in the form of text or speech, lexical coverage revealed from the data varies according to producer- and genre-related factors. Each individual collection of a corpus is only representative of the specific producer group under a given condition of language production. The Sinica Corpus is a balanced corpus of texts containing different genres. In the design of the TMC Corpus, we have attempted to cover varieties of formal and informal speaking situations by the arrangement of conversation partners (strangers vs. familiar persons) and different speaking scenarios by the arrangements of tasks (free conversation, map task, and topic-specific). It is clear that the TMC Corpus and the Sinica Corpus are not directly and completely comparable in terms of producers and genres. Nevertheless, the TMC Corpus and the Sinica Corpus were compiled by adopting the same word segmentation and POS tagging system, and they are currently the largest conversational and textual corpora available for Taiwan Mandarin. For this reason, when we examine the lexical coverage of the TMC Corpus, the Sinica Corpus will be compared to explore the similarities and differences among words produced in the form of conversation and text. Wordlists derived from these two corpora were used, the *Taiwan Mandarin Spoken*

[1] The *Taiwan Mandarin Spoken Wordlist* has been publicly distributed and can be freely downloaded from the website http://mmc.sinica.edu.tw/resources_e_01.htm.

Wordlist and the *Word List with Accumulated Word Frequency in Sinica Corpus 3.0* (CKIP, 1998). In order to collect information about syllables as well, we ran the same automatic conversion program to the *Word List with Accumulated Word Frequency in Sinica Corpus 3.0.* The results, however, were not manually checked, as we did for the TMC Corpus with the homograph errors.

Table 3. Conversational and Text Corpus.

Corpus	Word tokens	Word types	Syllable tokens	Syllable types with tones	Syllable types without tones
TMC Corpus	405,435	16,683	607,008	1,076	390
Sinica Corpus	4,767,048	55,301	7,515,036	1,120	392

For the current study, we have cleaned up errors we found in the wordlist of the Sinica Corpus, so the statistics summarized in **Table 3** may be slightly different from the official ones published by the CKIP team. As one can see, the Sinica Corpus is about ten times bigger than the TMC Corpus.

3.1 Word coverage

Corpus coverage of different vocabulary sizes in both corpora is listed in **Table 4**. The top 2000 word types in the TMC Corpus make up about 90% of the overall word tokens, whereas they only account for 70% of word tokens in the Sinica Corpus. McCarthy (1999: 236) has made a comparable proposal that "… a round-figure pedagogical target of the first 2000 words in order of frequency will safely cover the everyday core with some margin for error." Counting homographs with different POS categories as distinct word types, 1,117 among the top 2000 word types occur in both corpora, including nouns, verbs, adverbs, conjunctions, determinatives, prepositions, pronouns, non-predicative adjectives, particles, the structural particle DE, and the copula SHI. These 1,117 word types shared in the top 2000 list of both corpora eventually account for 81% of the TMC corpus coverage and 58% of the Sinica corpus coverage. A selection of these 1,117 word types, the (approximately) top 100 words in both corpora, is listed in **Appendix A**. They may be regarded as the core vocabulary that is required for operable communication in the form of conversation and text. For educational purposes, this core vocabulary may be the target words for teaching praxis and materials to focus on (Xiao *et al.*, 2009; Tao, 2009).

Table 4. Vocabulary Size and Corpus Coverage.

Vocabulary size	TMC corpus	Tokens	Tokens per type	Vocabulary size	Sinica corpus	Tokens	Tokens per type
1,000	84.43%	342,306	342	1,000	59.78%	2,935,763	2,936
2,000	89.87%	364,364	182	2,000	68.69%	3,373,419	1,687
3,000	92.53%	375,134	125	3,000	73.53%	3,610,951	1,204
4,000	94.20%	381,919	95	4,000	76.77%	3,770,449	943
5,000	95.34%	386,559	77	5,000	79.17%	3,887,992	778
6,000	96.21%	390,081	65	6,000	81.03%	3,979,351	663
7,000	96.95%	393,058	56	7,000	82.55%	4,054,042	579
8,000	97.44%	395,058	49	8,000	83.82%	4,116,688	515
9,000	97.94%	397,058	44	9,000	84.92%	4,170,368	463
10,000	98.35%	398,752	40	10,000	85.87%	4,217,103	422
	100%	**405,435**	**24**		**100%**	**4,767,048**	**86**

The word distribution in both corpora is presented in terms of the accumulative frequency in **Figure 2**. To achieve a 90% of corpus coverage, the first 15,000 frequency-ranked word types in the Sinica Corpus and the first 2,000 ones in the TMC Corpus are required. Calculating the proportions of these word types in their corpus share, 27% of the observed word types in the Sinica Corpus and 12% in the TMC Corpus would account for the majority of the lexical coverage of each corpus. This may suggest that these two different vocabulary sets are required for fluent communication in the form of text and conversation. The size of word types differs largely in both corpora, i.e. 15,000 versus 2,000. Nevertheless, if we view the number of characters involved in the two vocabulary sets, there are 2,964 different characters in the case of the Sinica Corpus and 1,065 in the TMC Corpus. A Chinese character is normally also a morpheme in Mandarin Chinese and is equivalent to a tone-specified syllable. The large number of homographs in Chinese leads to asymmetry between the number of tone-specified syllables from the phonological point of view and the number of characters from the orthographic point of view. The vocabulary sets required for a fluent communication above are equivalent to 1,065 tone-specified syllables for text (1,120 for the Sinica Corpus in total), and 654 for conversation (1,076 for the TMC Corpus in total).

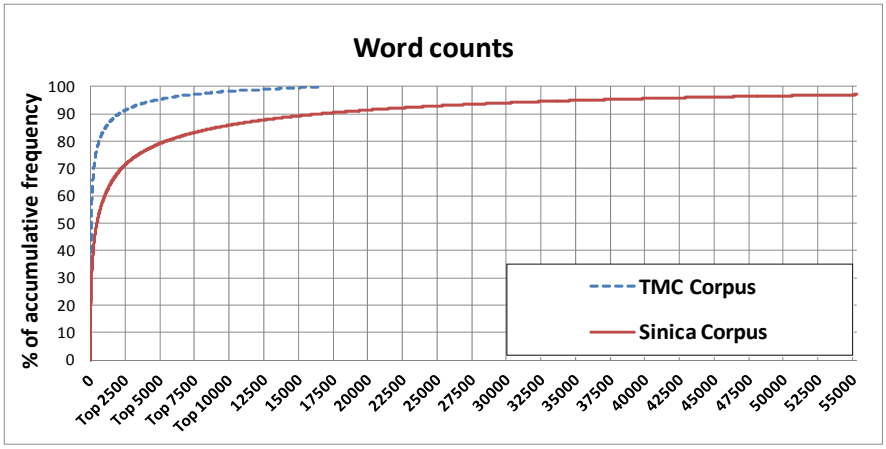


Figure 2. Word distribution.

3.2 Syllable Coverage

In the TMC Corpus, 1,076 different tone-specified syllable types were produced. In the Sinica Corpus, it was 1,120. Apparently, there is no clear difference between the text and conversation corpora in this regard, as shown in **Figure 3**. Similarly, to account for 90% of the corpus coverage, 300 tone-specified syllable types are required in the TMC Corpus and 400 are required in the Sinica Corpus. Moreover, if we disregard tone information, the number of syllable structures is 390 in the TMC Corpus, and 392 in the Sinica Corpus. This is almost the same in both corpora. Among them, 385 syllable structures were found in both corpora and the other 15 syllable structures appeared in only one of the corpora. The figures of syllables in both wordlists suggest that the capacity of phonologically different syllables (with or without considerations of lexical tones) in Taiwan Mandarin used in text and conversation is of similar size. Nevertheless, the number of tone-specified syllables does not equal the number of characters, or morphemes in Mandarin, as we mentioned earlier. For use in the form of text or conversation, the discrepancy is noticable, as the vocabulary sets required for fluent communication differ significantly: 1,065 for text and 654 for conversation.

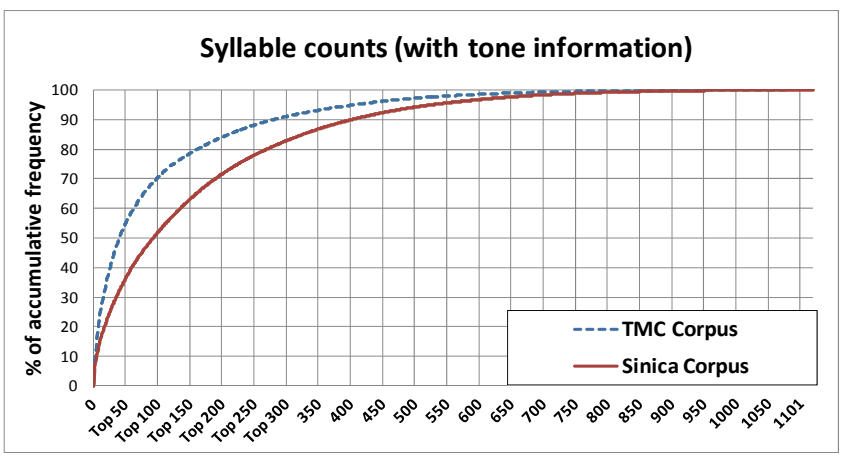


Figure 3. Syllable distribution.

3.3 Distribution of Word Category

The proportions of the 14 categories of the CKIP POS tags in both corpora are summarized in **Table 5**. The occurrences of nouns and verbs in the Sinica Corpus make up nearly 90% of the word tokens, suggesting that a certain percentage of nouns and verbs appear quite often in the Sinica Corpus. In contrast, the percentage of verbs and nouns in the TMC Corpus is only 45%. Words of the other categories, such as adverbs, pronouns, determinatives, prepositions, and conjunctions, were used significantly more often in conversation than in text.

Table 5. Word Category Distribution.

TMC Corpus	Coverage	Sinica Corpus	Coverage
Verb	23.30%	Noun	53.05%
Noun	22.05%	Verb	36.78%
Adverb	20.01%	Adverb	3.01%
Pronoun	9.98%	Determinative	2.60%
Determinative	6.42%	Foreign words	2.28%
Preposition	5.19%	Adjective	1.24%
Conjunction	4.66%	Conjunction	0.35%
Others	8.39%	Others	0.69%

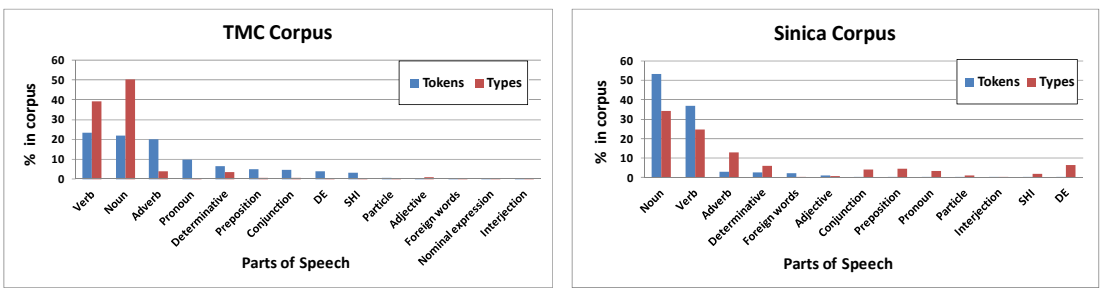


Figure 4. Parts of speech in conversational and text corpus.

The tokens per type of verb and noun in the Sinica Corpus are high because the corpus share of tokens is high and that of types is rather low, as shown in **Figure 4**. This may be due to the topics and the types of the articles included in the corpus, as the Sinica Corpus contains a large number of literary texts. On the contrary, the other word categories cover a much lower share of tokens, but more of types. In the TMC Corpus, a complementary distribution was observed. Verbs and nouns account for wider corpus coverage in terms of types than in terms of tokens. This suggests that different tasks and scenarios of conversations may elicit different

vocabularies. The other word categories, mostly function words, account for more tokens than types. In particular, the use of adverbs is different in conversation and in text. This, to a certain degree, is similar to the distribution found in a comparative study of spoken and written corpora of Swedish (Allwood, 1998). Adverbs, like the other function word categories (conjunctions and prepositions) were used more frequently in the spoken corpus than in the written corpus. Nevertheless, unlike in Taiwan Mandarin, pronouns and verbs were the most frequently produced categories in Swedish text and spoken corpora. The reason may lie in the characteristic of Chinese syntax. Zero anaphora is an often observed phenomenon in Chinese sentences. Therefore, pronouns are often used for addressing people in an interactive communication situation, for instance in conversation. As observed in the comparison of text and conversation, pronouns only make up 0.18% of the overall word tokens in the Sinica Corpus, but 10% in the TMC Corpus.

3.4 Discourse-related Items in the TMC Corpus

Interaction in conversation is often marked by pragmatic indicators, such as prosodic prominence, or by the use of discourse items, such as particles or feedback items. In this regard, conversation clearly differs from text. This section is concerned with corpus coverage of discourse-related items in the TMC Corpus. Compared with ordinary words, discourse items were produced much more frequently. The proportion of the occurrences of ordinary words over those of the discourse items is approximately eight to one in the TMC Corpus. That is, on average, a speech stretch of a length of eight words is accompanied by at least one discourse item. These items mark discourse-relevant positions in conversation, and they usually are produced with distinctive prosodic patterns to indicate the structure of a spoken discourse. With regard to information delivery, they may be considered a kind of redundancy. Their main function is to express the attitudes (particles), the fluency (markers and fillers), and the attention (feedback words) of the speakers. Without these discourse-related items, a conversation would be more like a scripted dialogue.

Table 6. Discourse-related items in conversation.

Groups	Tokens	Types	Tokens per type
Discourse particles	34,842	49	711
Discourse markers	16,516	9	1,835
Fillers/feedback words	6,338	65	98

For academic purposes, we need to investigate these discourse items, because they function as a kind of juncture between concepts and also function as markers of emerging patterns in conversation. As listed in **Table 6**, the tokens per type of discourse markers are 1,835, which is very high compared with ordinary words in the corpus. This suggests that the

performance of automatic speech recognition systems working with conversation can be improved in an economical and efficient way by implementing information and knowledge about the position of these discourse-related items (syntactic or prosodic) and their phonetic representation. Discourse particles are produced more often than the top 1000 word types in the TMC Corpus, 342 tokens per type. The numbers of the distinct types of discourse particles and markers are small, but the tokens per type are high. Furthermore, fillers and feedback words have a limited number of phonetic variants, as their phonetic representations are systematically predictable. Thus, they can be studied in terms of their phonetic forms, pronunciation variations, and their relationship to the contextual information. Feedback words normally mark the structure of speaker turn changes. Automatic detection of the discourse items would significantly enhance the understanding of conversation content and structure.

4. Conclusion

Spoken language is performed differently, given different speaking situations. To understand the lexical capacity of language users, no matter what purposes we have in mind, we need to base our investigations on realistic language data. The ideal corpus of this kind should take into account the versatility of speaker groups, conversation types, and speaking situations. In other words, it needs to be balanced among a variety of sociolinguistic settings. The concept of a balanced corpus for texts needs modification to be used for speech, as a balanced corpus of spoken data should also involve the spontaneous and interactive behavior of the speakers in specific speaking situations. Furthermore, the processing and presentation of speech corpora go beyond the consideration of the meta-data structures of text corpora. The transcribing convention needs to deal with the diversity of spoken phenomena in spontaneous speech. The alignment with the speech signal needs to manually or automatically be conducted to increase the innovative values of speech corpora applications for language technology system and language teaching tools. It is unlikely that the study of lexical coverage based on the Taiwan Mandarin Conversational Corpus represents the capacity of all Taiwan Mandarin speakers in all kinds of speaking situations. Nevertheless, we presented an attempt to provide empirical data for this line of research. With this data, we hope to extend our understanding about the notion how and why humans are capable of conversing by words for communication.

Acknowledgements

The author is grateful to the useful comments provided by two anonymous reviewers of the *International Journal of Computational Linguistics and Chinese Language Processing*. The author also sincerely thanks to the team members who have been working on the corpus data along the years. The study presented in this article is funded by the National Digital Archives Project and the National Science Council under Grant NSC-100-2410-H-001-093.

References

Allwood, J. (1998). Some frequency based differences between spoken and written Swedish. In *Proceedings of the XVIth Scandinavian Conference of Linguistics*, Department of Linguistics, University of Turku.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., & Weiner, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, 24(4), 351-366.

Baayan, R. H. (2001). *Word Frequency Distribution*. Kluwer Academic Publishers. Dordrecht/Boston/London.

Boersma, P. & Weenink, D. (2012). Praat: doing phonetics by computer. http://www.fon.hum.uva/praat 5.3.16.

Brown, G. D. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavior Research Methods, Instruments, & Computers*, 16(6), 502-532.

Chao, Y. R. (1965). *A Spoken Grammar of Chinese*. University of California Press.

Chen, K.-J. & Huang, C.-R. (1996). The SINICA CORPUS": Design methodology for balanced corpora. In *Proceedings of the Eleventh Pacific Asia Conference on Language, Information and Computation*, 167-176.

CKIP. (1998). *The Sinica Corpus 3.0*. The Chinese Knowledge Information Processing Group - technical report 98-04. Academia Sinica. (In Chinese)

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing*, 8(4), 259-265.

French, N., Carter, C. W., & Koenig, W. (1930). The words and sounds of telephone conversations. *Bell System Tech Journal*, 9, 290-324.

Gibbon, D., Moore, R., & Winski, R. (Eds.) (1997). *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter.

Howes, D. (1964). Application of the word frequency concept to aphasia. In A. V. S. de Reuck and M. O'Connor, *Disorders of Language* (Ciba Foundation Symposium). London: Churchill, 47-75.

Howes, D. (1966). A word count of spoken English. *Journal of Verbal Learning and Verbal Behavior*, 5(6), 572-606.

Knowles, G. (1990). The use of spoken and written corpora in the teaching of language and linguistics. *Literary and Linguistic Computing*, 5(1), 45-48.

Leech, G., Rayson, P., & Wilson, A. (2001). Word Frequencies in Written and Spoken English - Based on the British National Corpus. Pearson Education Limited.

McCarthy, M. (1999). What constitutes a basic vocabulary for spoken communication? *Studies in English Language and Literature*, 1, 233-249.

Reppen, R. & Ide, N. (2004). The American National Corpus: Overall Goals and the First Release. *Journal of English Linguistics*, 32, 105-113.

Schriffrin, D. (1988). *Discourse Markers*. Cambridge University Press.

Shriberg, E. E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. Doctoral dissertation. Department of Psychology, University of California at Berkeley.

Svartvik, J. & Quirk, R. (1980). *A Corpus of English Conversation*. Lund, Sweden: Gleerup.

Tao, H.-Y. (2009). Core Vocabulary in Spoken Mandarin and the Integration of Corpus-Based Findings into Language Pedagogy. In *Proceedings of the 21st North American Conference on Chinese Linguistics (NACCL-21)*. Vol. 1. Edited by Yun Xiao, 13-27. Smithfield, Rhode Island: Bryant University.

Thorndike, E.L. (1921). The Teacher's Word B*ook*. New York: Teachers College. Columbia University.

Thorndike, E.L. & Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University.

Tseng, S.-C. (2004). Processing spoken Mandarin corpora. *Traitement automatique des langues*. Special issue: Spoken corpus processing, 45, 89-108.

Xiao, R., Rayson, P., & McEnery, T. (2009). *A frequency dictionary of Mandarin Chinese: Core vocabulary for learners*. Routledge Frequency Dictionaries. London and New York: Taylor and Francis Group.

Wepman, J. M. & Lozar, B. (1973). The most frequently used words of spoken English. *Journal of Psycholinguistic Research*, 2(2), 129-136.

Appendix A: The top 100 words in the core vocabulary

Word	POS	TMC tokens	TMC %	Sinica tokens	Sinica %	Word	POS	TMC tokens	TMC %	Sinica tokens	Sinica %
的	DE	15778	3.89	28582	6.00	上	Ng	1339	0.33	8650	0.18
是	SHI	13999	3.45	84014	1.76	可	D	1337	0.33	8508	0.18
一	Neu	13397	3.30	58388	1.22	爲	VG	1300	0.32	8369	0.18
在	P	7429	1.83	56769	1.19	或	Caa	1296	0.32	8317	0.17
有	V_2	7092	1.75	45823	0.96	好	VH	1273	0.31	8304	0.17
個	Nf	6991	1.72	41077	0.86	等	Cab	1264	0.31	8070	0.17
我	Nh	6705	1.65	40332	0.85	又	D	1197	0.30	8037	0.17
不	D	6677	1.65	39014	0.82	將	D	1161	0.29	7858	0.16
這	Nep	6330	1.56	33659	0.71	後	Ng	1160	0.29	7752	0.16
了	Di	5453	1.34	31873	0.67	因爲	Cbb	1115	0.28	7592	0.16
他	Nh	5301	1.31	30025	0.63	於	P	1030	0.25	7395	0.16
也	D	5260	1.30	29646	0.62	由	P	1001	0.25	7344	0.15
就	D	4827	1.19	29211	0.61	從	P	989	0.24	7303	0.15
人	Na	4694	1.16	24269	0.51	更	D	971	0.24	7298	0.15
都	D	4473	1.10	20403	0.43	被	P	953	0.24	7272	0.15
說	VE	4419	1.09	19625	0.41	才	Da	877	0.22	7266	0.15
而	Cbb	4414	1.09	18452	0.39	已	D	863	0.21	7256	0.15
我們	Nh	4242	1.05	18152	0.38	者	Na	850	0.21	7221	0.15
你	Nh	4100	1.01	17298	0.36	每	Nes	841	0.21	7207	0.15
了	T	3882	0.96	15958	0.33	次	Nf	840	0.21	7087	0.15
要	D	3435	0.85	15955	0.33	把	P	837	0.21	7024	0.15
之	DE	3412	0.84	15893	0.33	三	Neu	834	0.21	6954	0.15
會	D	3398	0.84	14066	0.30	什麼	Nep	832	0.21	6729	0.14
對	P	3173	0.78	13944	0.29	問題	Na	814	0.20	6683	0.14
及	Caa	3124	0.77	13758	0.29	其	Nep	801	0.20	6667	0.14
和	Caa	2932	0.72	13585	0.28	讓	VL	782	0.19	6624	0.14
與	Caa	2832	0.70	13445	0.28	此	Nep	748	0.18	6599	0.14
以	P	2276	0.56	13172	0.28	做	VC	721	0.18	6597	0.14
很	Dfa	2189	0.54	13013	0.27	再	D	716	0.18	6563	0.14
種	Nf	2088	0.52	12263	0.26	所以	Cbb	708	0.17	6529	0.14

中	Ng	2066	0.51	12231	0.26	只	Da	684	0.17	6521	0.14
的	T	1976	0.49	11580	0.24	與	P	665	0.16	6519	0.14
大	VH	1926	0.48	11577	0.24	沒有	VJ	651	0.16	6510	0.14
能	D	1907	0.47	11125	0.23	則	D	646	0.16	6476	0.14
著	Di	1901	0.47	11026	0.23	台灣	Nc	633	0.16	6414	0.13
她	Nh	1869	0.46	10776	0.23	卻	D	630	0.16	6388	0.13
那	Nep	1848	0.46	10740	0.23	地	DE	620	0.15	6329	0.13
上	Ncd	1768	0.44	10619	0.22	並	Cbb	618	0.15	6171	0.13
但	Cbb	1697	0.42	10242	0.21	位	Nf	615	0.15	6015	0.13
年	Nf	1650	0.41	10127	0.21	得	DE	609	0.15	5969	0.13
還	D	1644	0.41	9698	0.20	去	D	604	0.15	5748	0.12
可以	D	1641	0.40	9671	0.20	呢	T	593	0.15	5577	0.12
時	Ng	1633	0.40	9565	0.20	學生	Na	593	0.15	5523	0.12
最	Dfa	1628	0.40	9416	0.20	表示	VE	592	0.15	5504	0.12
自己	Nh	1579	0.39	9069	0.19	到	P	572	0.14	5468	0.11
爲	P	1573	0.39	9026	0.19	公司	Nc	569	0.14	5421	0.11
來	D	1566	0.39	8992	0.19	將	P	568	0.14	5365	0.11
所	D	1518	0.37	8873	0.19	如果	Cbb	563	0.14	5336	0.11
他們	Nh	1500	0.37	8818	0.18	社會	Na	563	0.14	5282	0.11
各	Nes	1454	0.36	8651	0.18	看	VC	562	0.14	5198	0.11

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 1, March 2013, pp. 19-46

Learning to Find Translations and Transliterations on the Web based on Conditional Random Fields

Joseph Z. Chang[*], Jason S. Chang[+], and Jyh-Shing Roger Jang[#]

Abstract

In recent years, state-of-the-art cross-linguistic systems have been based on parallel corpora. Nevertheless, it is difficult at times to find translations of a certain technical term or named entity even with a very large parallel corpora. In this paper, we present a new method for learning to find translations on the Web for a given term. In our approach, we use a small set of terms and translations to obtain mixed-code snippets returned by a search engine. We then automatically annotate the data with translation tags, automatically generate features to augment the tagged data, and automatically train a conditional random fields model for identifying translations. At runtime, we obtain mixed-code webpages containing the given term and run the model to extract translations as output. Preliminary experiments and evaluation results show our method cleanly combines various features, resulting in a system that outperforms previous works.

Keywords: Machine Translation, Cross-lingual Information Extraction, Wikipedia, Conditional Random Fields.

1. Introduction

The phrase translation problem is critical to many cross-language tasks, including statistical machine translation, cross-lingual information retrieval, and multilingual terminology (Bian & Chen, 2000; Kupiec, 1993). Such systems typically use a bilingual lexicon or a parallel corpus to obtain phrase translations. Nevertheless, the out of vocabulary problem (OOV) is difficult to overcome, even with a very large training corpus, due to the Zipf nature of word

[*] Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, Taiwan
E-mail: joseph.nthu.tw@gmail.com
The author for correspondence is Joseph Z. Chang.
[+] Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
E-mail: jason.jschang@gmail.com
[#] Department of Computer Sciences and Information Engineering, National Taiwan University, Taiwan
E-mail: jang@csie.ntu.edu.tw

distribution and the fact that new words, technical terms, and named entities arise frequently. On the other hand, the advent of the Internet has lead to an unprecedented buildup of multilingual texts. Specifically, there are an abundance of webpages consisting of mixed-code text, namely text written in more than one language. We observe that the mixed-code webpages typically are written in one language but interspersed with some sentential or phrasal translations written in another language. By retrieving and identifying such translation counterparts on the Web, we can cope with the OOV problem caused by the limited coverage of dictionaries and parallel corpora.

Consider a Wikipedia title, "*Named-entity recognition*". The best places to find the Chinese translations for this technical term are probably not some parallel corpus or dictionary, but rather mixed-code webpages that mention it in both Chinese and English. The following example is a snippet returned by the *Bing* search engine for the query "*named entity recognition*" requesting Chinese language webpages:

> http://zh.wikipedia.org/zh-hk/問答系統: 從系統內部來看,問答系統使用了大量有別於傳統資訊檢索系統自然語言處理技術，如自然語言剖析(Natural Language Parsing)、問題分類(Question Classification)、專名辨識(Named Entity Recognition)等等。

In this snippets, the author mentioned several technical terms in Chinese (*e.g.*, 自然語言剖析 *zhiran yuyan poxi*, 問題分類 *wenti fenlei*, and 專名辨識 *zhuanming bianshi*), followed by the source terms in brackets (*Natural Language Parsing*, *Question Classification*, *and Named Entity Recognition*, respectively). The term-translation pairs in the above example follow the parenthetical translation surface pattern in the form of "*Chinese translation (English term)*". This pattern is only one of many surface patterns found on the Web that may indicate a term-translation pair. In the following examples, we show different surface patterns of translation pairs found on the Web, with Chinese translations underlined and the counterpart English terms italicized:

(a) 血液學檢驗(*hematology*) – 白血球分類

(b) [巴黎最美的橋] 亞歷山大三世橋 *Pont Alexandre III*

(c) 胰島素泵的臨床應用及護理進展 progress on nursing of clinical application of *insulin pump*

(d) 國外組織美國職棒大聯盟 (*Major League Baseball*，簡稱: *MLB*，或大聯盟)

(e) [食記]義美蔥油餅 Imei *green onion pancake*

(f) [食記]義美蔥油餅 *Imei* green onion pancake . . .

Examples (a) and (b) show Chinese translations occurring near or next to an English phrase. There are also cases (*e.g.*, Example (c)) where the translation (*e.g.*,胰島素泵 yidaoshu pang) and the English phrase (*e.g., insulin pump*) are far apart. Example (d) shows another form of parenthetical translation pattern, where translations are right next to the English term (*Major League Baseball*). Examples (e) and (f) show two term translation pairs interwoven in the same text (義美 yi-mei transliterated into *Imei* and 蔥油餅 cong-you-bing translated into *green onion pancake*).

For a given English term, such translations can be extracted by classifying the Chinese characters in the snippets as either translation or otherwise. Intuitively, we can cast the problem as a sequence labeling task. To be effective, we need to associate each token (*i.e.*, Chinese character or word) with some features to characterize the likelihood of the token being part of the translation. For example, by exploiting some external knowledge sources (*e.g.*, bilingual dictionaries), we derive that the Chinese character "辨" (*bian*) in the Chinese word "辨識" (*bian-shi, recognition*) is likely to be part of the translation of "*named entity recognition.*"

In this paper, we present a new method that automatically obtains such labeled data and generates features for training a conditional random fields (CRF) model that is capable of identifying translation or transliteration in mixed-code snippets returned by search engines (*e.g., Google* or *Bing*). The system uses a small set of phrase-translation pairs to obtain search engine snippets that may contain both an English term and its Chinese translation from search engines. The snippets then are tagged automatically to train a CRF sequence labeler. We describe the training process in more detail in Section 4.

At run-time, we start with a given phrase (*e.g.*, "*named-entity recognition*"), which is transformed into a query with a setup to retrieve webpages in the target language (*e.g.*, Chinese). We then retrieve mixed-code snippets returned by the search engine and extract translations within the snippets. The identified translations can be used to supplement a bilingual terminology bank (*e.g.*, adding multilingual titles to existing Wikipedia); alternatively, they can be used as additional training data for a machine translation system, as described in Lin, Zhao, Van Durme, and Paşca (2008).

Most previous works focus on extracting translation pairs where the counterpart terms appear near one another in the webpage, based on a limited set of short patterns. In our approach, we extract term and translation pairs that are near or far apart, and are not limited by a set of predefined patterns. We have evaluated our method based on English-Chinese

language links in Wikipedia as the gold standard. Results show that our method produces output for 80% of the test cases with an exact match precision of 43%, outperforming previous works.

The rest of the paper is organized as follows. In the next Section 2, we survey the related work that also aimed to mine translations from the Web. In Section 3, we give brief descriptions on resources we make use of. In Section 4, we describe in detail the problem statement and the proposed method. Finally, we report evaluation results and error analysis in Section 5.

2. Related Work

In machine translation, a source text is typically translated one sentence at a time, while cross-lingual information retrieval involves phrasal translation. The proposed methods for phrase translation in the literature rely on either handcrafted bilingual dictionaries, transliteration tables, or bilingual corpora. For example, Knight and Graehl (1998) described and evaluated a multi-stage machine translation method for performing backwards transliteration of Japanese names and technical terms into English, while Bian and Chen (2000) described cross-language information access to multilingual collections on the Internet. Recently, Smadja, McKeown, and Hatzivassiloglou (1996) proposed an algorithm for producing collocation and translation pairs, including noun and verb phrases, in bilingual corpora. Similarly, Kupiec (1993) propose an algorithm for finding noun phrase correspondence in bilingual corpora for bilingual lexicography and machine translation. Koehn and Knight (2003) described a noun phrase translation subsystem that improves word-based statistical machine translation methods.

Some methods in the literature also have aimed to exploit mixed code webpages for word and phrase translation. Nagata, Saito, and Suzuki (2001) presented a system for finding English translations for a given Japanese technical term in search engine results. Their method extracts English phrases appearing near the given Japanese term, and it scores translation candidates based on co-occurrence counts and location. Cao and Li (2002) proposed an EM algorithm for finding translation for base noun phrases on the Web. Kwok *et al*. (2005) focused on named entity phrases and implemented a cross-lingual name finder based on Chinese-English webpages. Wu, Lin, and Chang (2005) proposed a method for learning a set of surface patterns to find terms and translations occurring in short distance. Mixed-code webpage snippets were obtained by querying a search engine with English terms for Chinese webpages. They discovered that the most frequent pattern is where the translation immediately followed by the source term, with the coverage rate of 46%. Their results also indicate the stricter parenthetical pattern covers less than 30% of the translation instances.

Researchers also have explored the hyperlinks in webpages as a source of bilingual

information. Lu, Chien, and Lee (2004) proposed a method for mining terms and translations from anchor text directly or transitively. In a follow-up project, Cheng *et al*. (2004) proposed a method for translating unknown queries with web corpora for cross-language information retrieval. Similarly, Gravano and Henzinger (2006) also proposed systems and methods for using anchor text as parallel corpora for cross-language information retrieval.

In a study more closely related to our work, Lin *et al.* (2008) proposed a method that performs word alignment between Chinese translations and English phrases within parentheses in crawled webpages. Their paper also proposed a novel and automatic evaluation method based on Wikipedia. The main difference from our work is that the alignment process in Lin *et al*. (2008) is done heuristically using a competitive linking algorithm proposed by Melamed (2000), while we use a learning-based approach to align words and phrases. Moreover, in their method, only *parenthetical translations* are considered. With only the parenthetical pattern, their method is able to extract a significant number of translation pairs from crawled webpages without a given list of target English phrases. By restricting to parenthetical surface patterns however, many translation pairs in webpages may not be captured, including term-translation pairs that are further apart. In our work, we exploit surface patterns differently as a soft constraint in a CRF model and use an approach similar to Lin *et al*. (2008) to evaluate our results.

In contrast to the previous work in phrase and query translation, we present a learning-based approach that uses annotated data to develop the system. Nevertheless, we do not require human intervention to prepare the training data, but instead make use of language links in Wikipedia to automatically obtain the training data. The annotated data is further augmented with features indicative of translation and transliteration relations obtained from external lexical knowledge sources publicly-available on the Web. The trained CRF sequence labeler then is used to find translations on the Web for a given term.

3. Resources

In this work, we rely on several resources that are available on the Internet. These resources are used for different purposes: the seed data are used for obtaining and labeling training data, the gold standard is used for automatic evaluation, and the external knowledge sources are used for generating features.

3.1 Wikipedia

Wikipedia is an online encyclopedia compiled by volunteers around the world. Anyone on the Internet can edit existing entries or create new entries to add to Wikipedia. Owing to the number of its participants, Wikipedia has achieved both high quantity and a quality comparable to traditional encyclopedias compiled by experts (Giles, 2005). Due to these

reasons, Wikipedia has become the largest and most popular reference tool.

We extracted bilingual title pairs from the English and Chinese editions of Wikipedia as the gold standard for evaluation and as seeds to automatically collect and label training data from the Internet by querying search engines.

The number of entries in English Wikipedia grew at an exponential rate from 2001 to 2008, with some 20,000 new articles created monthly by thousands of volunteers around the world, making it an excellent source for finding new words and terms. As of February 2, 2012, the English Wikipedia had 3,861,652 articles, making it the most well-established edition for all 285 languages.

Entries on the same topic among different language editions of Wikipedia are interlinked via the so-called language links. Nevertheless, only a small percentage of English articles are linked to editions of other languages. The Chinese Wikipedia contains only 398,206 articles, making it roughly one-tenth the size of the English Wikipedia. Furthermore, only 5% of the entries in the English Wikipedia contain language links to their Chinese counterparts. The proposed method can be used to find the translations of those English terms, thus speeding up the process of building a more complete multilingual Wikipedia. As will be described in Section 4, we extracted the titles of English-Chinese article pairs connected by language links for training and testing purposes.

The content of Wikipedia is freely downloadable online.[1] We used the *Google Freebase Wikipedia Extraction* (*WEX*) instead of the official raw dump. The *WEX* is a processed version of the official dump, with the Wikipedia syntax transformed into XML. The *WEX* database can be freely downloaded online.[2]

3.2 WordNet

WordNet is a freely available, handcrafted lexical semantic database for English.[3] Starting its development in 1985 at Princeton University by a team of cognition scientists, WordNet was originally intended to support psycho-linguistic research. Over the years, WordNet has become increasingly popular in the fields of information retrieval, natural language processing, and artificial intelligent. Through each release, WordNet has grown into a comprehensive database of concepts in the English language. As of today, the stable 3.0 version of WordNet contains 207,000 semantic relations between 150,000 words organized in over 115,000 senses.

Senses inWordNet are represented as synonym sets (*synsets*). A synset with a definition contains one or more words, or *lemmas*, that express the same meaning. In addition, WordNet

[1] http://en.wikipedia.org/wiki/Wikipedia:Database_download

[2] http://wiki.freebase.com/wiki/WEX

[3] http://wordnet.princeton.edu/

provides other information for each synset, including example sentences and estimated frequency. For example, the synset {*block, city_block*} is defined as *a rectangular area in a city surrounded by streets*, whereas synset {*block, cube*} is defined as *a three-dimensional shape with six square or rectangular sides*. WordNet also records various semantic relations between its senses. These relations includes *hypernyms*, *hyponyms*, *coordinate terms*, *holonym* and *meronym*.

3.3 Sinica Bilingual WordNet

The *Sinica Bilingual WordNet* is part of the publicly accessible *Sinica Bilingual Ontological WordNet* (*Sinica BOW*) (Huang, 2003). In this work, we treat the *Sinica Bilingual WordNet* as a bilingual dictionary, and use it as an external knowledge source to generate features for training the CRF model.

The Sinica Bilingual WordNet is a hand-crafted English-Chinese version of the original Princeton WordNet 1.6. It was compiled by collecting all possible Chinese translations of a synset's lemmas from various online bilingual dictionaries before a team of translators manually edited the acquired translations. For each synset, the translators selected at most three appropriate lexicalized words as translation equivalents.

The *Sinica BOW* system can be freely-accessible online.[4] The *Sinica Bilingual WordNet* database can also be licensed for download.[5]

3.4 NICT Bilingual Technical Term Database

The *NICT Bilingual Technical Term Database* is a resource freely available online.[6] In addition to the *Sinica Bilingual WordNet*, we also used the NICT database to generate features. While the Sinica Bilingual WordNet mainly contains common nouns, the NICT database mainly contains technical terms and proper nouns. By combining the two resources, we can generate translational features covering both common nouns and proper nouns.

The NICT Bilingual Technical Term Database is maintained by committees in the National Academy for Educational Research of Taiwan (formerly National Institute for Compilation and Translation). The goal is to pursuit more uniform and standardized translations for technical terms used in textbooks, patents, national standards, and open source software. It contains over 1.1 million Chinese-English term translation pairs arranged into 72 categories (Table 9) and is kept up to date by constantly including more terms. Any user can suggest a new term and translation to the committees to be added to the database.

---

[4] http://BOW.sinica.edu.tw/

[5] http://www.aclclp.org.tw/doc/bw_agr_e.PDF

[6] http://terms.nict.gov.tw/

3.5 Google Web 1T N-grams

In 2006, *Google* published a ngram dataset based on public webpages through Linguistics Data Consortium for licensing.[7] The Google Web 1T corpus is a 24 GB (gzip compressed) corpus that consists of n-grams ranging from unigram to five-grams generated from approximately 1 trillion words in publicly accessible Web pages. In this work, we use the Web 1T corpus to filter unlinked entries in the English Wikipedia with high frequency on the Web for manual evaluation.

4. Method

Submitting an English phrase (*e.g.*, "*named-entity recognition*") to search engines to find translations or transliteration is a good strategy used by many translators (Quah, 2006). Unfortunately, the user has to sift through snippets to find the translations. Such translations usually exhibit characteristics related to word translation, word transliteration, surface patterns, and proximity to the occurrences of the given phrase. To find translations for a given term on the Web, a promising approach is automatically learning to extract phrasal translations or transliterations of a given query using the conditional random fields (CRF) model. To avoid human effort in preparing annotated data for training the model, we use an automatic procedure to retrieve and tag mixed-code search engine snippets using a set of bilingual Wikipedia titles. We also propose using external knowledge sources (*i.e.*, bilingual dictionaries, name lists and terminology banks) to generate translational and transliterational features.

4.1 Problem Statement

We focus on the issue of finding translations in mixed code snippets returned by a search engine. The translations are identified, tallied, ranked, and returned as the output of the system. The returned translations can be used to supplement existing multilingual terminology banks, or used as additional training data for a machine translation system. Therefore, our goal is to return several reasonably precise translations that are available on the Web for the given phrase.

Problem Statement: Given a phrasal term *P* and a full-text search engine *SE* (*e.g.*, *Bing* or *Google*) that operates over a mixed-code document collection (*e.g.*, the Web), our goal is to retrieve a probable translation *T* of *P* via *SE*.

For this, we extract a set of translation candidates, c_1, ..., c_m from a set of mixed-code snippets, s_1, ..., s_n returned by *SE*, such that these candidates are likely to be translations *T* of *P*.

[7] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

(1) Retrieve mixed-code snippets and tag translations (Section 4.3.1)
(2) Generate translation features (Section 4.3.2)
(3) Generate transliteration features (Section 4.3.3)
(4) Generate distance features (Section 4.3.4)
(5) Train a CRF model for classifying translations (Section 4.3.4)

Figure 1. Outline of the training phase.

In the rest of this section, we describe our solution to this problem. First, we briefly introduce the Conditional Random Fields (CRF) model in Section 4.2. We describe a strategy (see Figure 1) for obtaining training data for identifying translation in snippets returned by *SE* (Section 4.3.2). This strategy relies on a set of term-translation pairs for training, derived from Wikipedia language links (Section 4.3.1). We will also describe our method for exploiting external knowledge sources to generate translation features (Section 4.3.2), transliteration features (Section 4.3.3), and distance features (Section 4.3.4) for sequence labeling. Finally, in Section 4.4, we describe how to extract and filter translations at run-time by applying the trained sequence labeler.

4.2 Conditional Random Fields

Sequence labeling is the task of assigning labels from a finite set of categories to a sequence of observations. This problem is encountered in the field of computational linguistics, as well as in many other fields, including bio-informatics, speech recognition, and pattern recognition.

Traditionally, the sequence labeling problem are often solved using the Hidden Markov Model (HMM) or Maximum Entropy Markov Model (MEMM). Both HMM and MEMM are directed graph models in which every outcome is conditioned on the corresponding observation node and the previous outcomes (*i.e.*, Markov property).

Conditional Random Fields (CRF), proposed by Lafferty, McCallum, and Pereira (2001), is considered the state-of-the-art sequence labeling algorithm. One of the major differences of CRF is that it is modeled as an undirected graph. For sequence labeling, the CRF graph is structured as an undirected linear chain (chained CRF). CRF obeys the Markov property with respect to the undirected graph, as every outcome is conditioned on its neighboring outcomes and potentially the entire observation sequence. In our case, the outcomes are B, I, O labels that indicate a sequence of Chinese characters in the search engine snippets that is likely the translation or transliteration of the given English term. The information available (the observable) for sequence labeling are the characters in the snippets themselves, and the three types of features we generate.

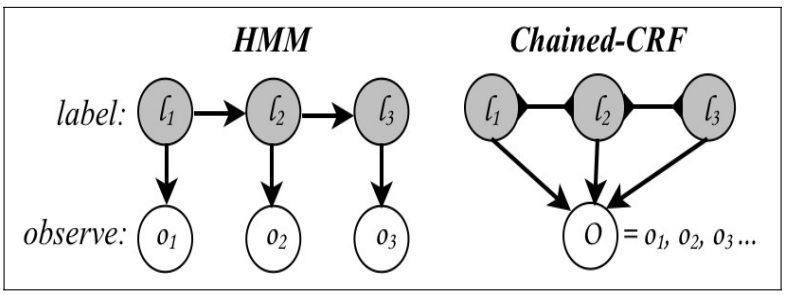


Figure 2. Simplified view of HMM and CRF.

4.3 Preparing Data for CRF Classifier

We attempt to learn to find translations or transliterations for given phrases on the Web. For this, we make use of language links in Wikipedia to obtain seed data, retrieve mixed-code snippets returned by a search engine, and augment feature values based on external knowledge sources. Our learning process is shown in Figure 1.

4.3.1 Retrieving and Tagging Snippets

In the first stage of the training phase, we extracted Wikipedia English titles and their Chinese counterparts using the language links as the seed data for training. We use the English titles to query a search engine (*e.g.*, *Google* or *Bing*) with the target Web page language set to Chinese. This strategy will bias the search engine to return Chinese web pages interspersed with some English phrases. We then automatically labeled each Chinese character in the returned snippets, using the common *BIO* notation, with *B*, *I*, *O* indicating the beginning, inside, and outside of translations, respectively (*e.g.*, 支援向量機 *zhiyuan-xiangliang-ji*). An additional *E* tag is used to indicate the occurrences of the given term (*e.g.*, *support vector machine*).

1. …1995/O 年/O 提/O 出/O 的/O **支/B 持/I 向/I 量/I 機/I** (/O *support/E vector/E machine/E*，/O SVM/O)/O 以/O 訓/O 練/O …
2. …發/O 光/O 原/O 理/O 不/O 同/O。/O **光/B 通/I 量/I** *luminous/E flux/E* 光/O 源/O 在/O 單/O 位/O 時/O 間/O …

***Figure 3. Examples of tagged snippets for title pairs* "support vector machine", "支持向量機" *and* "luminous flux", "光通量".**

The output of this stage is a set of tagged snippets that can be used to train a statistical sequence classifier for identifying translations. A sample of two tagged snippets, automatically generated from bilingual Wikipedia titles are shown in Figure 3. The *E* tags are designed to provide proximity cues for labeling the translation and capture common surface patterns of the phrase and translation in mixed code data. For example, in Figure 3, the translation 支持向量機(*zhichi xiangliang ji*) is tagged with one B tag and four I tags,

followed by the left parenthesis and three *E* tags. The translation 光通量 (*guangtong liang*) is tagged with one *B* tag and two *I* tags, immediately followed by two *E* tags. Such sequences (*i.e. B I I I I O E E E*, and *B I I E E*) are two of many common patterns.

Note that we do not attempt to produce word alignment information, as done in Lin *et al.* (2008). In contrast, we only use the BIO labeling scheme to indicate phrasal translations, leading to a smaller number of parameters required to be estimated during the training process.

4.3.2 Generating Translation Features

We generate translation features using external bilingual resources with the ϕ^2 score proposed by Gale and Church (1991) to measure the correlations between an English word and a Chinese character:

$$\phi^2 = \frac{[P(e,f)P(\bar{e},\bar{f}) - P(\bar{e},f)P(e,\bar{f})]^2}{P(e)P(f)P(\bar{e})P(\bar{f})} \quad (1)$$

where *e* is an English word and *f* is a Chinese character occurring in bilingual phrase pairs.

Table 1. Example of a Chinese-English dictionary with three entries.

Chinese	English
社交工程	social engineering
社群網路	social network
社群媒體	social media

Table 2. Example of English word and Chinese character probability.

w	Count(w)	P(w)	P($\bar{w}$)	e	f	Count(e,f)	P(e,f)
社	3	1.00	0.00	social	社	3	1.00
群	2	0.67	0.33	social	群	2	0.67
交	1	0.33	0.67	social	交	1	0.33
網	1	0.33	0.67	network	社	1	0.33
social	3	1.00	0.00	network	群	1	0.33
media	1	0.33	0.67	network	交	0	0.00
network	1	0.33	0.67	network	網	1	0.33

In our case, the ϕ^2 scores are calculated by counting the occurrence of Chinese characters and English words in the publicly-available bilingual dictionaries or termbanks. To illustrate, we use a tiny Chinese-English dictionary in Table 1 with only three entries to explain how the probabilities are calculated. We treat each entry in the dictionary as an event, and calculate the probability of each Chinese character and English word by counting the

number of events containing them, as shown in Table 2. Similarly, we can calculate the joint probability of an English word and a Chinese character by counting their co-occurrences in the dictionary.

Table 3. Three contingency tables indicating co-occurrence and none co-occurrence.

	vector	
向	793	9,960
	97	1,975,642

	vector	
量	768	21,907
	122	1,963,695

	machine	
機	3,381	28,566
	491	1,954,054

In Table 3, we show the contingency table calculated by counting co-occurrences in Bilingual WordNet and NICT termbank for (向 *xiang*, *vector*), (量 *liang*, *vector*), and (機 *ji*, *machine*). The statistical association between an English word (*e.g.*, *vector*) and its translation (*e.g.*, 向 (*xiang*)) is indicated by the high count of co-occurrences, as well as the lower values of two inverse diagonal cells. From the contingency tables, we can calculate the corresponding ϕ^2 scores for 向 *xiang* 量 *liang*, and 機 *ji*: 0.06530, 0.02880, and 0.09068.

Table 4. Example ϕ^2 scores.

	support	vector	machine
提	0.00000	0.00000	0.00000
出	0.00000	0.00000	0.00000
的	0.00000	0.00000	0.00000
支	**0.09075**	0.00000	0.00000
持	**0.00058**	0.00000	0.00000
向	0.00000	**0.06530**	0.00000
量	0.00000	**0.02880**	0.00000
機	0.00000	0.00000	**0.09067**

	luminous	flux
發	**0.00432**	0.00000
光	**0.01028**	6.0E-06
原	0.00000	0.00000
理	0.00000	0.00000
不	1.4E-06	0.00000
光	**0.01028**	6.0E-06
通	0.00000	**0.06410**
量	0.00000	**0.00793**

To generate features for each token, we calculate the following logarithmic value of ϕ^2:

$$feat_{translation}(f) = 9 + log(\underset{e \in E}{argmax}\phi^2(e, f)) \quad (2)$$

where *e* is a word in the given English phrase *E*, and *f* is the Chinese character in a snippet. This feature value is rounded to a whole number in order to limit the number of distinct feature values. In Table 4, we show the ϕ^2 scores of each Chinese character in snippets from searching Google with the given terms, *i.e.*, *support vector machine* and *luminous flux*. Notice that there are some noisy feature values in the second example: the Chinese characters in the word 發光 (*faguang*, *glow* or *illuminate*) has non-zero ϕ^2 scores. However, the tagger potentially can overcome such noise by relying on other features, such as the distance feature (Section 4.3.4). Moreover, in most cases there are multiple snippets for a given term, from which we can confidently identify the translations with higher frequencies. As an example, we

show two snippets tagged with translation features in Figure 4. In this example, the translation characters are given feature values ranging from 2 to 7, while non-translation ones are mostly 0.

1. ... 1995/0 年/0 提/0 出/0 的/0 **支/7 持/2 向/6 量/5 機/7** (/0 *support*/*E* *vector*/*E machine*/*E*，/0 SVM/0)/0 以/0 訓/0 練/0 ...
2. ... 發/0 光/5 原/0 理/0 不/0 同/0 。/0 **光/5 通/7 量/5** *luminous*/*E* *flux*/*E* 光/5 源/0 在/0 單/0 位/0 時/0 間/0 ...

***Figure 4. Example of two snippets tagged with translation features given the terms* "support vector machine" *and* "luminous flux".**

4.3.3 Generating Transliteration Features

We generate the additional features related to transliteration using some external knowledge resources. It is important to include transliteration in the feature set, since many named entities or technical terms are transliterated in full or partially into a foreign language. Thus, the translation feature described in Section 4.3.2 alone is not enough. For this, we collect transliterated titles from the entries connected with language links across the English and the Chinese Wikipedia to calculate correlation between the target transliteration characters and English sublexical strings.

We observed that names of persons and geographic locations are mostly transliterated, and that the entries titled with names of persons or locations can be extracted easily from Wikipedia using the categories of each entry. As will be described in Section 5, we extracted Wikipedia articles tagged with categories that match "*Birth in ...*" to find articles describing a person, and categories that matches "*Cities in ...*" and "*Capitals in ...*" to find titles describing a geographic location. We show some named entities in Table 6.

Table 5. English words segmentation for Chinese-English syllable alignment.

Chinese Transliteration	Chinese Romanization	English Named Entity	Possible Segmentations
喬布斯	qiao-bu-si	jobs	j-o-bs, j-ob-s, **jo-b-s**
瓊喬	qiong-qiao	jonjo	j-onjo, jo-njo, **jon-jo**, jonj-o
喬瑟夫	qiao-se-fu	joseph	j-o-seph, j-os-eph, j-ose-ph, j-osep-h, jo-s-eph, **jo-se-ph**, jo-sep-h, jos-e-ph, ...
喬凡尼	qiao-fan-ni	giovanni	g-i-ovanni, g-io-vanni, g-iov-anni, ..., gio-va-nni, **gio-van-ni**, gio-vann-i, ...

Table 6. Force alignment results of Chinese and English transliteration examples.

Chinese Transliteration	Chinese Romanization	English Syllables
喬布斯	qiao-bu-si	jo-b-s
瓊喬	qiong-qiao	jon-jo
喬瑟夫	qiao-se-fu	jo-se-ph
喬凡尼	qiao-fan-ni	gio-van-ni
拉喬利納	la-qiao-li-na	ra-joe-li-na
奧喬亞	ao-qiao-ya	o-cho-a

After obtaining the transliteration pairs from Wikipedia, we align the Chinese and English syllables. In Chinese, every character always represents one syllable. Nevertheless, the counterpart "syllables" in an English word are not as easy to determine. These counterparts are not syllables in the regular sense, for some counterpart "syllables" may contain a single consonant. We assume every extracted Chinese and English transliteration pairs contain the same number of syllables, *i.e*., equal to the number of Chinese characters. We also assume the syllables are transliterated in order. Under these assumptions, we can segment the English words into a number of segments equal to the number of characters in its Chinese transliteration, and align the English segments and Chinese characters in order. For example, as shown in Table 5, the English name *Joseph* is transliterated into three Chinese characters, or syllables, 喬瑟夫 *qiao-se-fu*, therefore, all possible segmentations include: *j-o-seph*, *j-os-eph*, *j-ose-ph*, *j-osep-h*, *jo-s-eph*, *jo-se-ph*, *jo-sep-h*, *jos-e-ph*, ..., etc.

We use the Expectation-Maximization (EM) algorithm to estimate the conditional probabilities $P(f|e)$ modeling the correlation between the Romanized Chinese characters and the English counterpart. For Chinese characters that have ambiguous pronunciations, we use the Romanization of the most frequent pronunciation according to the Chinese Electronic Dictionary from Academia Sinica, available for download from the The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).[8] In the E-step, the expectation of the log-likelihood of each segmentation candidates are evaluated using the current estimation of $P(f|e)$. In the M-step, the conditional probability estimations are updated based on the maximum likelihood estimation (MLE) of the E-step. A few examples of the segmentation results are shown in Table 6.

[8] http://www.aclclp.org.tw/use_ced.php

Table 7. Conditional probability of Chinese Romanized Chinese character with English syllable. Note that many Chinese characters typically shared the same Romanization.

Rom. Chinese	English Tr.	Cnt(f,e)	P(f|e)
qiao	geo	140	0.38
	jo	66	0.18
	joe	41	0.11
bu	b	1090	0.58
	bu	301	0.16
	br	122	0.07
si	s	5626	0.69
	es	292	0.04
	st	226	0.03

After aligning the syllables in the transliteration pairs, we then calculate the conditional probability of the Romanized Chinese character and its English counterpart. Example output of three Romanized Chinese characters and their top English counterparts is shown in Table 7.

Nevertheless, generating transliteration features for each Chinese character (Romanized) tends to produce a lot of false positives. Therefore, we assume that a named entity is transliterated into at least two Chinese characters, and generate the transliteration features of a Chinese character taking into consideration the preceding and following characters. Admittedly, we probably missed some transliteration cases, such as *Jean* and 琴 (*qin*), but that represents a small loss.

In general, this strategy works quite well for our purpose. For example, given the character sequence 喬布斯(*qiao-bu-si*) and the term *Steve Jobs*, to calculate the transliteration score for the Chinese character 布(*bu*), we calculate the probability of 喬布(*qiao-bu*) and 布斯(*bu-si)* being part of transliteration of *Steve* or *Jobs*:

$$\begin{aligned} P(bu \mid steve) &= max(P(qiao-bu \mid steve), P(bu-si \mid steve)) \\ P(bu \mid jobs) &= max(P(qiao-bu \mid jobs), P(bu-si \mid jobs)) \end{aligned} \tag{3}$$

To calculate the conditional probability for the Chinese bi-characters 喬布 *qiao-bu* given the English term *jobs*, we generate all substring *xy* of *jobs*, into which *qiao-bu* can be transliterated:

$$\begin{aligned} &P(qiao-bu \mid jobs) = \underset{xy \in jobs}{argmax}(P(qiao \mid x) \mid (bu \mid y)) \\ &xy \in jobs \text{ denotes string } xy \text{ is a substring of } jobs \end{aligned} \tag{4}$$

With this probabilistic value, we then generate the transliteration feature values in a similar way as described in Section 4.3.2:

$$feat_{transliteration}(f) = 9 + log(\underset{e \in E}{argmax}\, P(f \mid e)) \quad (5)$$

1. ... 法-fa/0 國-guo/0 立-li/0 體-ti/2 主-zhu/2 義-yi/0 畫-hua/0 家-jia/4 **喬-qiao/7 治-zhi/7 ・/0 布-bu/8 拉-la/8 克-ke/4** (/0 *georges/E braque/E*)/0 ...
2. ... 第-di/0 62/0 屆-jie/0 **艾-ai/3 美-mei/3 獎-jiang/0** 頒-ban/0 獎-jiang/0 典-dian/0 禮-li/0 》/0(/0 the/0 62nd/0 *Emmy/E Award/E*)/0 ...

Figure 5. Example of transliteration features given* Georges Braque *to find the Chinese transliteration "*喬治・布拉克*" and given* Emmy Award *to find "*艾美獎*"

We show two examples of the data tagged with transliteration feature values in Figure 5. In the first example, given the phrase *Georges Braque*, the name of a French painter, to find its Chinese transliteration "喬治・布拉克 (*qiao-zhi bu-la-ke*)". The respective feature scores for each of the characters in the transliteration are *7 7 0 8 8 4*. The symbol "・" with a feature value of zero, is commonly used in Chinese name transliteration to identify the boundary of first and last name in foreign names, and it can be identified as part of the answer by its surrounding transliteration feature scores and the surface pattern. Also in the first example, the Chinese character 家(*jia*), the second syllable of 畫家(*hua-jia*, painter), has a noisy non-zero feature value of *four*, due to the fact that the English syllable *geo* is often transliterated into this Chinese syllable *jia*. In the second example, the given phrase is *Emmy Award*, where the first part of the phrase *Emmy* is transliterated into 艾美(*ai-mei*), and the second part of the phrase *Award* is translated in to 獎(*jiang*). The Chinese characters 艾 and 美 both have a feature value of 3, while all other characters in the example have a feature value of zero. We also show this example tagged with all types of feature values we generate in Table 8.

4.3.4 Generating Distance Features

Finally, we generate the distance features and train a CRF model. The distance feature is intended to exploit the fact that translations tend to occur near the source term, as pointed out in Nagata *et al*. (2001) and Wu *et al*. (2005). Therefore, we incorporated the distance as an additional feature type, to impose a soft constraint on the locational relations between a translation and its English counterpart.

An example showing all three kinds of features and labels is shown in Table 8. This example shows that the given term *Emmy Award* has a Chinese counterpart that is part transliteration (*Emmy* with a transliteration 艾美 *ai-mei*) and part translation (*Award* with the translation 獎 *jiang*). This is a typical case that our method is designed to handle using both

translational and transliterational features. Finally, we use the labeled data with three kinds features to train a CRF model.

Table 8. Example training data.

	word	TR	TL	distance	label
	第	0	0	14	O
	62	0	0	13	O
(62nd)	屆	0	0	12	O
	艾	3	0	11	B
(Emmy)	美	3	0	10	I
(Award)	獎	0	5	9	I
	頒	0	0	8	O
(awarding)	獎	0	0	7	O
	典	0	0	6	O
(ceremony)	禮	0	0	5	O
	》	0	0	4	O
	(	0	0	3	O
	the	0	0	2	O
	62nd	0	0	1	O
	Emmy	0	0	0	E
	Award	0	0	0	E
	)	0	0	-1	O

4.4 Runtime Translation Extraction

Once the CRF model is automatically trained, we attempt to find translations for a given phrase using the procedure in Figure 6.

In Step 1, the system submit the given phrase as query to a search engine (*SE*) to retrieve snippets. Then, for each token in each snippet, we generate three kinds of features (Step 2). This process is exactly the same as in the training phase. In Step 3, we run the CRF model on the snippets to generate labels. Then, in Step 4, we extract the Chinese strings with a sequence of *B*, *I*, ..., *I* tags as translation candidates.

Finally, in Step 5, we compute the frequency of all of the candidates identified in all snippets, and output the candidate with the highest frequency as output. When there is a tie

with multiple candidates with the same highest frequency, one of them is randomly selected as the output.

```
Procedure FindTranslation(P, SE):

(1)Submit P as a query to SE
       to retrieve a set of mixed-code snippets s1, s2, s3, ..., sn

   for each snippet si in snippets s1, s2, s3, ..., sn:
        for each Chinese character in si:
(2)         Generate the three features base on P

(3)Run the CRF model on snippets with features for BIO labels

   for each snippet si in snippets s1, s2, s3, ..., sn:
(4)    Extract Chinese tagged with BI sequence as candidates

(5)Output the candidate with highest redundancy (frequency).
  (In case of a tie, randomly select one of the most frequent.)
```

Figure 6. Pseudocode of the runtime phase.

5. Evaluation

We extracted the titles of English and Chinese articles that are connected through language links in Wikipedia using the Wikipedia dump created on 2010/08/16 (Google, 2010). We used a short list of stop words based on the rules pointed out by Lin *et al*. (2008) to exclude titles that are for administrative or other purposes. We obtained a total of 155,310 article pairs, from which we randomly selected 13,150 and 2,181 titles as seeds to obtain the training and test data, respectively, as described in Section 4.3.1. We then used the English-Chinese Bilingual WordNet[9] and NICT terminology bank (terms.nict.gov.tw/download_main.php) to generate translational features, in an effort to cover both common nouns and technical terms. The bilingual WordNet, translated from the original Princeton WordNet 1.6 has 99,642 synset entries, each with multiple lemmas and multiple translations, forming a total of some 850,000 translation pairs. The NICT database has over 1.1 million term translation pairs in 72 categories and covers a wide variety of different fields. See Table 9 for the numbers of entries in each of the 72 categories.

[9] http://www.aclclp.org.tw/doc/bw_agr_e.PDF

Table 9. Categories of the NICT term database.

Category	Count	Category	Count
Pharmacy	1,673	Material Science (Polymer)	3,422
Bacterial Immunology	2,063	Material Science (Ceramics)	2,292
Phylogenetic	1,756	Agricultural Machinery	3,060
Psychopathology	1,067	Science Education	5,289
Psychology	5,741	Industrial Engineering	5,400
Physics/Chemistry Equipments	17,279	Astronomy	6,091
Comparative Anatomy	6,013	Music	2,922
Education	2,198	Food Science and Technology	35,666
Sociology	2,825	Foreign Names	57,054
Human Anatomy	5,796	Mineralogy	28,032
Pathology	7,307	Lab Animal and Comparative Medicine	8,220
Sports	1,708	Dance	10,564
Soil Science	1,240	Statistic	7,370
Forestry	7,954	Meteorology	20,061
Fertilizer Science	1,155	Animal Husbandry	21,466
Hydraulic Engineering	4,601	Mining and Metallurgical Engineering	13,914
Electronic Engineering	7,627	Computer	101,389
Agricultural Promotion	669	Textile Science and Technology	2,2761
Accounting	4,884	Meteorology	17,789
Civil Engineering	16,745	Endocrinology	2,577
Aeronautics and Astronautics	23,751	Chemical Engineering	22,386
Electrical Engineering	20,058	Communications Engineering	16,899
Engineering Graphics	4,766	Biology (Plants)	42,730
Mathematics	16,708	Mechanism and Machine Theory	2,085
Foundry	5,314	Shipbuilding Engineering	30,701
Mechanical Engineering	35369	Physics	22,077
Earth Science	30673	Zoology	29,586
Geology	22780	Marine	37,329
Marketing	1667	Chemistry (Compound)	19,258
Veterinary Medicine	24,990	Fish	29,730
Nuclear Energy	38,462	Economics	8,891
Production Automation	2,560	Marine Geology	31,015
Surveying	14,371	Power Engineering	69,546
Ecology	7,495	Chemistry (Others)	25,273
Mechanics	10,716	Administration	3,743
Materials Science (Metal)	7,665	Journalism and Communication	4,419

For transliterational features, we extracted person or location entries in Wikipedia using such categories as "*Birth in ...*" to find titles for a person, and categories such as "*Cities in ...*" and "*Capitals in ...*" to find titles for a geographic location. A total of some 15,000 bilingual person names and 24,000 bilingual place names were obtained and forced aligned.

To compare our method with previous work, we used a similar evaluation procedure as described in Lin *et al*. (2008). We ran the system and produced the translations for these 2,181 test data, and we automatically evaluated the results using the metrics of coverage and exact match precision based on the Wikipedia language links. We removed all search snippets from the *wikipedia.org* domain to ensure a strict separation of training and test datasets.

This precision rate is an underestimation since a term may have many alternative translations that do not match exactly with the single reference translation. To obtain a more accurate estimate of the real precision rate, we resorted to manual evaluation.

We selected a small part of the 2,181 English phrases and manually evaluated the results. We report the results of automatic evaluation in Section 5.1 and the results of manual evaluation in Section 5.2.

5.1 Automatic Evaluation

In this section, we describe the evaluation based on the set of 2,181 English-Chinese title pairs extracted from Wikipedia as the gold standard and automatically evaluate coverage (applicability) and exact match precision. Coverage is measured by the percentage of titles for which the proposed system produces some translations.

When translations were extracted, we selected the most frequent translations as output, and checked for exact match against the reference answer. Table 10 shows the results we obtained as compared to the results reported by Lin *et al.* (2008).

We explored the performance differences of the systems employing different set of features. The systems evaluated are as follows:

- **Full**: the proposed system trained with all feature types.
- **-TL** : the proposed system trained without the transliteration feature.
- **-TR** : the proposed system trained without the translation feature.
- **-TL-TR** : the proposed system only using the distance feature. No external knowledge used.
- **LIN En-Ch** : the results reported in the Lin *et al*. paper for their system targeting Chinese parenthetical translations.
- **LIN** En-Ch : the results reported in the Lin *et al*. paper for their system targeting English parenthetical translations

Table 10. Automatic evaluation results.

system	coverage	exact match	top5 exact match
Full (En-Ch)	80.4%	43.0%	56.4%
-TL	83.9%	27.5%	40.2%
-TR	81.2%	37.4%	50.3%
-TL-TR	83.2%	21.1%	32.8%
LIN En-Ch	59.6%	27.9%	not reported
LIN Ch-En	70.8%	36.4%	not reported
LDC (En-Ch)	10.8%	4.8%	N/A
NICT (En-Ch)	24.2%	32.1%	N/A

• **LDC** : the LDC2.0 English to Chinese bilingual dictionary with 161,117 translation pairs. (reported in Lin *et al.*)

• **NICT** : the freely available NICT technical term bilingual dictionary with 1,138,653 translation pairs.

Notice that, although Lin *et al.* (2008) also used bilingual Wikipedia title pairs for evaluation, they used an earlier snapshot of Wikipedia and worked with full webpages crawled from the Internet without a list of given terms. We worked with the list of English terms given as input, but worked only with search engine snippets. In the previous work, all of the bilingual title pairs extracted from Wikipedia were used for evaluation. In our work, only a portion of the title pairs were used for evaluation and the rest were used for generating the training data. It is often difficult to compare systems with different experimental settings. Nevertheless, the evaluation results seem to indicate that the proposed method compares favorably with the results reported in the previous work.

With a given target English term as input, the proposed system uses a search engine to retrieve a relevant portion of limited webpages, and attempts to find the Chinese translation within the retrieved text. The proposed system extracts translations in all cases without being limited by a set of a few surface patterns, and has a significantly higher coverage and precision rate than the previous method that rely on the parenthetic patterns only.

As shown in Table 10, we found using external knowledge to generate features improves system performance significantly. Adding translation feature (-TL) or transliteration feature (-TR) improves exact match precision by about 6% and 16%, respectively. Due to the fact that many Wikipedia titles are fully or partially transliterated into Chinese, the transliteration feature was found to be more important than the translation feature.

The results also clearly show that finding translations on the Web has the advantage of

better coverage than simply looking up phrases in a terminology bank (with a coverage rate of 24%), or a bilingual dictionary (with a coverage rate of 11%). Although using the NICT terminology bank or LDC bilingual dictionary directly has the worst performance, using them as external knowledge sources improves the performance of the CRF model significantly.

Overall, the full system performed the best, finding translations for 8 out of 10 phrases with an average exact match precision rate of over 40%. Nearly 60% of the exact matches appear in the Top 5 candidates. Leaving out the transliteration feature degraded the precision rate by 16%, far more than leaving out the translation feature. This is to be expected, since English Wikipedia has considerably more named entities with transliterated counterparts in Chinese.

5.2 Manual Evaluation

In this section, we present two sets of manual evaluation. In Section 5.2.1, we manually evaluate the results produced by the full system.

5.2.1 Error Analysis on Automatic Evaluation

Since an English phrase is often translated into several Chinese counterparts, evaluation based on exact match against a single reference answer leads to under-estimation. Therefore, we asked a human judge to examine and mark the output of our full system. The judge was instructed to mark each output as **A**: correct translation alternative, **B**: correct translation but with a difference sense from the reference, **P**: partially correct translation, and **E**: incorrect translation.

Table 11 shows 24 randomly selected translations that do not match the relevant reference translations. Half of the translations (12) are correct translations (**A** and **B**), while a third (8) are partially correct translation (**P**). Notice that it is a common practice to translate only the surname of a foreign person. So, four of the eight partial translations may be considered as correct.

In Table 12, we show extracted candidates and frequency counts for 8 example terms. Translation candidates are marked using the same *A*, *B*, *P*, and *E* tags as in Table 11, plus an additional tag, *M*, to indicate an exact match. For the given term *money laundering*, the system extracted 27 exact matches (洗錢), and 2 correct alternatives (洗黑錢) and only 1 erroneous output from 30 snippets returned from the search engine. While technical terms like *money laundering* tend to have literal translations and result in more exact matches, movie titles are often translated into Chinese with completely different meanings. For example, the official Chinese title for the movie, *Music and Lyrics* in Taiwan is “K-歌-情人” (meaning *karaoke-song-lover*). Given such a title as input, the system was able to extract 18 partial

matches and 2 exact matches base on surface patterns and modest translation feature value for *music* and 歌(*ge*, *song*). For the given term *colony*, the system extracted 菌落(*colony of fungi or bacteria*), a correct translation with a different sense. Other extracted answers include: transliteration, 科羅尼海島酒店(*Island Colony*), the name of a hotel, and the exact-match translation, 殖民地(*foreign control territory*). For the given term *bubble sort*, the partial translation 排序(*sort*) makes the top-1 translation (with a count of 20), while the top-2 to top-5 are either exact-match or acceptable translations.

Table 11. Cases failing the exact match test.

English Wiki	Chinese Wiki	Extracted	
Pope Celestine IV	塞萊斯廷四世	切萊斯廷四世	A
Huaneng Power International	華能國際	華能國際電力	A
Shangrao	上饒市	上饒	A
Aurora University	震旦大學	奧羅拉大學	A
Fujian	福建省	福建	A
Dream Theater	夢劇場	夢劇場合唱團	A
Coturnix	鶉屬	鵪鶉	A
Waste	垃圾	廢物	A
Allyl alcohol	烯丙醇	丙烯醇	A
Machine	機械	工具機	A
Colony	殖民地	菌落	B
Collateral	落日殺神	抵押	B
Ludwig Erhard	路德維希・艾哈德	艾哈德	P
John Woo	吳宇森	約翰	P
Osman I	奧斯曼一世	奧斯曼	P
Itumeleng Khune	伊圖梅倫・庫內	庫內	P
Naphthoquinone			P
Base analog	鹼基類似物	鹼基類	P
Chinese Paladin	仙劍奇俠傳	神劍	P
Bubble sort	冒泡排序	排序	P
The Love Suicides at Sonezaki	曾根崎情死	夏目漱石	E
Survivor's Law II	律政新人王II	金石良緣	E
Phichit	批集府	朗家庭主婦	E
Ammonium	銨	過硫酸銨	E

Note that this learning-based approach to mining translation and transliteration on the Web is an original contribution of our work. Previous works such as Wu *et al*. (2005); Lin *et al*. (2008), simply used occurrence statistics to identify translations, which is roughly equivalent to our translational or transliterational features (see Section 4.3.2 and Section 4.3.3). While Lin *et al*. used prefixes of 3 letters to provide a makeshift model of transliteration, we model the name-transliteration relations directly using an EM algorithm. Moreover, we also take note of their pattern of appearance to allow more effective extraction of relevant translations with the distance feature (see Section 4.3.4). It is important to note that combining features inherent in a training data, as well as derived from external knowledge sources in a machine learning model allow us to cover more relevant translations, while filtering out many invalid candidates.

Table 12. Extracted candidates and frequencies.

given term	freq	candidate	
money laundering	27	洗錢	**M**
	2	洗黑錢	A
	1	洗錢宣傳	E
Music and Lyrics	18	歌情人	P
	2	K歌情人	**M**
flyback transformer	14	變壓器	P
	3	回掃變壓器	**M**
	2	返馳式變壓器	A
	2	返馳變壓器	A
colony	15	菌落	B
	2	科羅尼海島酒店	B
	2	殖民地	**M**
Osman I	8	奧斯曼	P
	5	奧斯曼一世	**M**
bubble sort	20	排序	P
	19	泡排序	A
	17	氣泡排序	**M**
	9	泡沫排序	A
	4	泡泡排序	A

6. Conclusion and Future Work

We have presented a new method for mining translations on the Web for a given term. In our work, we use a set of terms and translations as seeds to obtain mixed-code snippets returned by a search engine, such as *Google* or *Bing*. We then automatically convert the snippets into a tagged sequence of tokens, automatically augment the data with features obtained from external knowledge sources, and automatically train a CRF model for sequence labels. At runtime, we submit a query consisting of the given term to a search engine, tag the returned snippets using the trained model, and finally extract and rank the translation candidates for output. Preliminary experiments and evaluations show our method cleanly combining various features, resulting in an integrated, learning-based system capable of finding both term translations and transliterations.

Many avenues exist for future research and improvement of our system. For example, existing query expansion methods to retrieve more webpages containing translation for the given phrases could be implemented (Zhang *et al*., 2005). Translation features related to word parts (*e.g*., *-lite* in the term *zeolite*) could be used to improve identification of translations. Additionally, an interesting direction to explore is to identify phrase types and length (*i.e*., base NP and NP prep. NP) and train type-specific CRF models for better results. In addition, natural language processing techniques such as word stemming, word lemmatization, or derivational morphological transformation could also be attempted to improve recall and precision.

Another interesting direction to explore is using a robot to crawl webpages and filter mixed-code data to derive the translation features. With the crawled web pages, we can extract translations offline, without having to work with a search engine and its limited returned snippets.

Yet another direction of research would be to enhance the effectiveness of translation features by working on the level of Chinese words instead of characters. For that, we could either use an existing, general-purpose word segmenter or carry out self-organized word segmentation (Sproat & Shih, 1990) to produce word-based translation features.

Reference

Bian, G.-W., & Chen, H.-H. (2000). Cross-language information access to multilingual collections on the internet. *Journal of the American Society for Information Science*, 51(3), 281-296.

Cao, Y., & Li, H. (2002). Base noun phrase translation using web data and the em algorithm. In *Proceedings of the 19th international conference on computational linguistics*, volume 1, 1-7.

Chang, J. Z., Chang, J. S., & Jang, R. J.-S. (2012). Learning to find translations and transliterations on the web. In *Proceedings of the 50th annual meeting of the association for computational linguistics*, volume 2, 130-134.

Cheng, P.-J., Teng, J.-W., Chen, R.-C., Wang, J.-H., Lu, W.-H., & Chien, L.-F. (2004). Translating unknown queries with web corpora for cross-language information retrieval. In *Proceedings of the 27th annual international acm sigir conference on research and development in information retrieval*, 146-153.

Fellbaum, C. (1998). *Wordnet: An electronic lexical database*. MIT Press.

Gale, W. A., & Church, K. W. (1991). Identifying word correspondence in parallel texts. In *Proceedings of the workshop on speech and natural language*, 152-157.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070), 900-901.

Google. (2010). Freebase data dumps (August 16th, 2010 ed.). http://download.freebase.com/datadumps/.

Gravano, L., & Henzinger, M. H. (2006). Systems and methods for using anchor text as parallel corpora for cross-language information retrieval (No. 7146358).

Huang, C.-R. (2003). Sinica bow: integrating bilingual wordnet and sumo ontology. In *Proceedings of 2003 International Conference on Natural Language Processing and Knowledge Engineering*, 825-826.

Huang, F., Vogel, S., & Waibel, A. (2003). Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In *Proceedings of the acl 2003 workshop on multilingual and mixed-language named entity recognition*, 15, 9-16.

Knight, K., & Graehl, J. (1998). Machine transliteration. *Computational Linguistics*, 24(4), 599-612.

Koehn, P., & Knight, K. (2003). Feature-rich statistical translation of noun phrases. In *Proceedings of the 41st annual meeting on association for computational linguistics*, volume 1, 311-318.

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st annual meeting on association for computational linguistics*, 17-22.

Kwok, K., Deng, P., Dinstl, N., Sun, H., Xu, W., Peng, P., & Doyon., J. (2005). Chinet: a chinese name finder system for document triage. In *Proceedings of 2005 international conference on intelligence analysis*.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning*, 282-289.

Li, Y., & Grefenstette, G. (2005). Translating chinese romanized name into Chinese idiographic characters via corpus and web validation. In *Proceedings of coria 2005*, 323-338.

Lin, D., Zhao, S., Van Durme, B., & Paşca, M. (2008). Mining parenthetical translations from the web by word alignment. In *Proceedings of acl-08: Hlt*, 994-1002.

Lu, W.-H., Chien, L.-F., & Lee, H.-J. (2004). Anchor text mining for translation of web queries: A transitive translation approach. *ACM Trans. Inf. Syst.*, 22(2), 242-269.

Melamed, I. D. (2000). Models of translational equivalence among words. *Computational Linguistics*, 26(2), 221-249.

Nagata, M., Saito, T., & Suzuki, K. (2001). Using the web as a bilingual dictionary. In *Proceedings of the workshop on data-driven methods in machine translation*, volume 14, 1-8.

Qu, Y., & Grefenstette, G. (2004). Finding ideographic representations of Japanese names written in latin script via language identification and corpus validation. In *Proceedings of the 42nd annual meeting on association for computational linguistics*.

Quah, C. K. (2006). *Translation and technology*. Palgrave Macmillan.

Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22(1), 1-38.

Sproat, R. W., & Shih, C. (1990). A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4(4), 336-351.

Wu, J.-C., Lin, T., & Chang, J. S. (2005). Learning source-target surface patterns for web-based terminology translation. In *Proceedings of the acl 2005 on interactive poster and demonstration sessions*, 37-40.

Zhang, Y., Huang, F., & Vogel, S. (2005). Mining translations of oov terms from the web through cross-lingual query expansion. In *Proceedings of the 28th annual international acm sigir conference on research and development in information retrieval*, 669-670.

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 1, March 2013, pp. 47-78

Machine Translation Approaches and Survey for Indian Languages

Antony P. J.*

Abstract

The term Machine Translation is a standard name for computerized systems responsible for the production of translations from one natural language into another with or without human assistance. It is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another. Many attempts are being made all over the world to develop machine translation systems for various languages using rule-based as well as statistically based approaches. Development of a full-fledged bilingual machine translation (MT) system for any two natural languages with limited electronic resources and tools is a challenging and demanding task. In order to achieve reasonable translation quality in open source tasks, corpus based machine translation approaches require large amounts of parallel corpora that are not always available, especially for less resourced language pairs. On the other hand, the rule-based machine translation process is extremely time consuming, difficult, and fails to analyze accurately a large corpus of unrestricted text. Even though there has been effort towards building English to Indian language and Indian language to Indian language translation system, unfortunately, we do not have an efficient translation system as of today. The literature shows that there have been many attempts in MT for English to Indian languages and Indian languages to Indian languages. At present, a number of government and private sector projects are working towards developing a full-fledged MT for Indian languages. This paper gives a brief description of the various approaches and major machine translation developments in India.

Keywords: Corpus, Computational Linguistics, Statistical Approach, Interlingua Approach, Dravidian Languages0.

* Professor and Head, Department of ISE, St. Joseph Engineering College, Mangalore, VTU.
E-mail: antonypjohn@gmail.com

1. Introduction

MT refers to the use of computers to automate some of the tasks or the entire task of translating between human languages. Development of a full-fledged bilingual MT system for any two natural languages with limited electronic resources and tools is a challenging and demanding task. Many attempts are being made all over the world to develop MT systems for various languages using rule-based as well as statistical-based approaches. MT systems can be designed either specifically for two particular languages, called a bilingual system, or for more than a single pair of languages, called a multilingual system. A bilingual system may be either unidirectional, from one Source Language (SL) into one Target Language (TL), or may be bidirectional. Multilingual systems are usually designed to be bidirectional, but most bilingual systems are unidirectional. MT methodologies are commonly categorized as direct, transfer, and Interlingua. The methodologies differ in the depth of analysis of the SL and the extent to which they attempt to reach a language independent representation of meaning or intent between the source and target languages. Barriers in good quality MT output can be attributed to ambiguity in natural languages. Ambiguity can be classified into two types: structural ambiguity and lexical ambiguity.

India is a linguistically rich area. It has 18 constitutional languages, which are written in 10 different scripts. Hindi is the official language of the Union. Many of the states have their own regional language, which is either Hindi or one of the other constitutional languages. In addition, English is very widely used for media, commerce, science and technology, and education only about 5% of the world's population speaks English as a first language. In such a situation, there is a large market for translation between English and the various Indian languages.

Even though MT in India started more than two decades ago, it is still an ongoing process. The third section of this paper discusses various approaches used in English to Indian languages and Indian language to Indian language MT systems. The fourth section gives a brief explanation of different MT attempts for English to Indian languages and Indian languages to Indian languages.

2. History of MT

The major changeovers in MT systems are as shown in Figure 1. The theory of MT pre-dates computers, with philosophers 'Leibniz and Descartes' ideas of using code to relate words between languages in the seventeenth century (Hutchins *et al*., 1993). The early 1930s saw the first patents for 'translating machines'. Georges Artsrouni was issued a patent in France in July 1933. He developed a device, which he called a *'cerveau mécanique'* (mechanical brain) that could translate between languages using four components: memory, a keyboard for input,

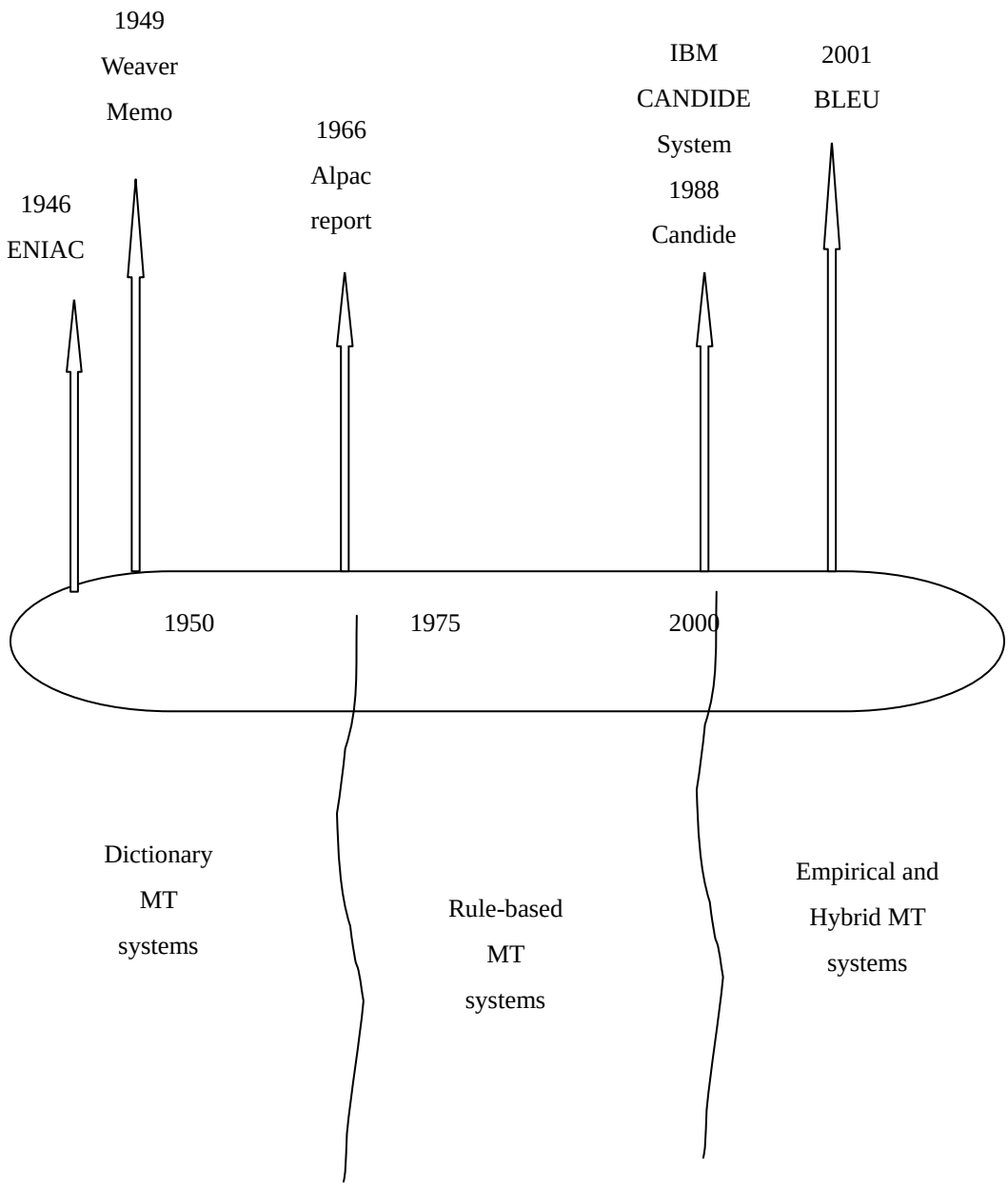


Figure 1. Major changeovers in MT Systems.

a search method, and an output mechanism. The search method was basically a dictionary look-up in the memory; therefore, Hutchins is reluctant to call it a translation system. The proposal Russian Petr Petrovich Troyanskii patented in September 1933 bears a resemblance

to the Apertium system, using a bilingual dictionary and a three-staged process, *i.e.* first a native speaking human editor of the SL (SL) pre-processed the text, then the machine performed the translation, and finally a native-speaking human editor of the TL post-edited the text (Hutchins *et al.,* 1993; Hutchins *et al.,* 2000).

After the birth of computers Electrical Numerical Integrator and Calculator (ENIAC) in 1947, research began on using computers as aids for translating natural languages (Hutchins *et al.,* 2005). The first public demonstration of MT in the Georgetown-IBM experiment, which proved deceptively promising, encouraged financing of further research in the field. In 1949, Weaver wrote a memorandum, putting forward various proposals (based on the wartime successes in code breaking) on the developments in information theory and speculation about universal principles underlying natural languages (Weaver *et al.,* 1999). In the decade of optimism, from 1954-1966, researchers encountered many predictions of imminent 'breakthroughs'. In 1966, the Automated Language Processing Advisory Committee (ALPAC) report was submitted, which said that, for 'semantic barriers', there are no straightforward solutions. The ALPAC report committee could not find any "pressing need for MT" nor "an unfulfilled need for translation (ALPAC *et al.,* 1996)".

This report brought MT research to its knees, suspending virtually all research in the United States of America (USA) while some research continued in Canada, France, and Germany (Hutchins *et al.,* 2005). After the ALPAC report, MT almost was ignored from 1966-1980. In the year 1988, Georgetown-IBM experiment launched "IBM CANDIDE System," where over 60 Russian sentences were translated smoothly into English using 6 rules and a bilingual dictionary consisting of 250 Russian words, with rule-signs assigned to words with more than one meaning. Although Professor Leon Dostert cautioned that this experimental demonstration was only a scientific sample, or "a Kitty Hawk of electronic translation (Kitty Hawk[1])," a wide variety of MT systems emerged after 1980 from various countries and research continued on more advanced methods and techniques. Those systems mostly were comprised of indirect translations or used an 'interlingua' as an intermediary. In the 1990s, Statistical Machine Translation (SMT) and what is now known as Example-based Machine Translation (EBMT) saw the light of day (IBM, 1954). At this time the focus of MT began to shift somewhat from pure research to practical application using a hybrid approach. Moving towards the change of the millennium, MT became more readily available to individuals via online services and software for their personal computers.

[1] Kitty Hawk, North Carolina, USA was the site for the world's first successful powered human flight by the Wright brothers. "Kitty Hawk" references generally meant a break-through success in its early stages.

3. MT Approaches

Generally, MT is classified into seven broad categories: rule-based, statistical-based, hybrid-based, example-based, knowledge-based, principle-based, and online interactive based methods. The first three MT approaches are the most widely used and earliest methods. Literature shows that there have been fruitful attempts using all these approaches for the development of English to Indian languages as well as Indian languages to Indian languages. At present, most of the MT related research is based on statistical and example-based approaches. Figure 2 shows the classification of MT in Natural language Processing (NLP).

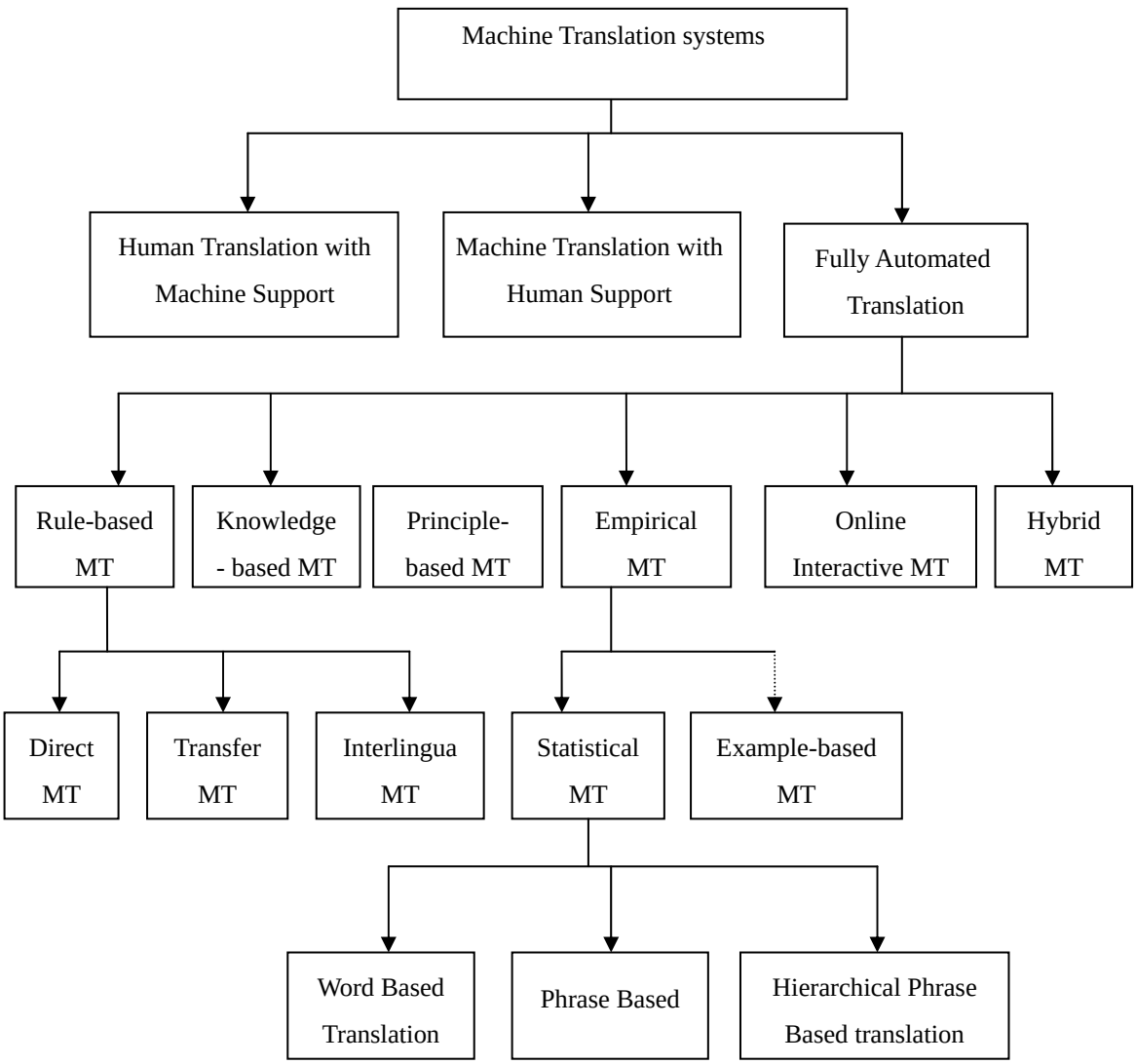


Figure 2. Classification of MT System.

3.1 Rule-based Approach

In the field of MT, the rule-based approach is the first strategy that was developed. A Rule-Based Machine Translation (RBMT) system consists of collection of rules, called grammar rules, a bilingual or multilingual lexicon, and software programs to process the rules.

Nevertheless, building RBMT systems entails a huge human effort to code all of the linguistic resources, such as source side part-of-speech taggers and syntactic parsers, bilingual dictionaries, source to target transliteration, TL morphological generator, structural transfer, and reordering rules. Nevertheless, a RBMT system always is extensible and maintainable. Rules play a major role in various stages of translation, such as syntactic processing, semantic interpretation, and contextual processing of language. Generally, rules are written with linguistic knowledge gathered from linguists. Transfer-based MT, Interlingua MT, and dictionary-based MT are the three different approaches that come under the RBMT category. In the case of English to Indian languages and Indian language to Indian language MT systems, there have been fruitful attempts with all four approaches. The main idea behind these rule-based approaches is as follows.

3.1.1 Direct Translation

In the direct translation method, the SL text is analysed structurally up to the morphological level and is designed for a specific source and target language pair (Noone *et al*., 2003; Dasgupta & Basu, 2008). The performance of a direct MT system depends on the quality and quantity of the source-target language dictionaries, morphological analysis, text processing software, and word-by-word translation with minor grammatical adjustments on word order and morphology.

3.1.2 Interlingua Based Translation

The next stage of progress in the development of MT systems is the Interlingua approach, where translation is performed by first representing the SL text into an intermediary (semantic) form called Interlingua. The advantage of this approach is that Interlingua is a language independent representation from which translations can be generated to different TLs. Thus, the translation consists of two stages, where the SL is first converted in to the Interlingua (IL) form before translation from the IL to the TL. The main advantage of this Interlingua approach is that the analyzer of the parser for the SL is independent of the generator for the TL. There are two main drawbacks in the Interlingua approach. The first disadvantage is, difficulty in defining the interlingua. The second disadvantage is Interlingua does not take the advantage of similarities between languages, such as translation between Dravidian languages. Nevertheless the advantage of Interlingua is it is economical in situations where translation among multiple languages is involved (Shachi *et al*., 2001).

Starting with the shallowest level at the bottom, direct transfer is made at the word level. Moving upward through syntactic and semantic transfer approaches, the translation occurs on representations of the source sentence structure and meaning, respectively. Finally, at the interlingual level, the notion of transfer is replaced with a single underlying representation called the 'Interlingua'. 'Interlingua' represents both the source and target texts simultaneously. Moving up the triangle reduces the amount of work required to traverse the gap between languages at the cost of increasing the required amount of analysis and synthesis.

3.1.3 Transfer Based Translation

Because of the disadvantage of the Interlingua approach, a better rule-based translation approach was discovered, called the transfer approach. Recently, many research groups have being using this third approach for their MT system, both abroad and in India. On the basis of the structural differences between the source and target language, a transfer system can be broken down into three different stages: i) Analysis, ii) Transfer and iii) Generation. In the first stage, the SL parser is used to produce the syntactic representation of a SL sentence. In the next stage, the result of the first stage is converted into equivalent TL-oriented representations. In the final step of this translation approach, a TL morphological analyzer is used to generate the final TL texts.

3.2 Statistical-based Approach

The statistical approach comes under Empirical Machine Translation (EMT) systems, which rely on large parallel aligned corpora. Statistical machine translation is a data-oriented statistical framework for translating text from one natural language to another based on the knowledge and statistical models extracted from bilingual corpora. In statistical-based MT, bilingual or multilingual textual corpora of the source and target language or languages are required. A supervised or unsupervised statistical machine learning algorithm is used to build statistical tables from the corpora, and this process is called the learning or training (Zhang *et al.*, 2006). The statistical tables consist of statistical information, such as the characteristics of well-formed sentences, and the correlation between the languages. During translation, the collected statistical information is used to find the best translation for the input sentences, and this translation step is called the decoding process. There are three different statistical approaches in MT, Word-based Translation, Phrase-based Translation, and Hierarchical phrase based model.

The idea behind SMT comes from information theory. A document is translated according to the probability distribution function indicated by $p(e|f)$, which is the Probability of translating a sentence f in the SL F (for example, English) to a sentence e in the TL E (for example, Kannada).

The problem of modeling the probability distribution *p(e|f)* has been approached in a number of ways. One intuitive approach is to apply Bayes theorem. That is, if *p(f|e)* and *p(e)* indicate translation model and language model, respectively, then the probability distribution $p(e|f) \propto p(f|e)p(e)$. The translation model *p(f|e)* is the probability that the source sentence is the translation of the target sentence or the way sentences in *E* get converted to sentences in *F*. The language model *p(e)* is the probability of seeing that TL string or the kind of sentences that are likely in the language *E*. This decomposition is attractive as it splits the problem into two sub problems. Finding the best translation $\tilde{e}$ is done by picking the one that gives the highest probability, as shown in Equation 1.

$$\tilde{e} = \arg\max_{e \in e^*} p(e \mid f) = \arg\max_{e \in e^*} p(f \mid e)\, p(e) \tag{1}$$

Even though phrase based models have emerged as the most successful method for SMT, they do not handle syntax in a natural way. Reordering of phrases during translation is typically managed by distortion models in SMT. Nevertheless, this reordering process is entirely unsatisfactory, especially for language pairs that differ a lot in terms of word-order. In the proposed project, the problem of structural differences between source and target languages is overcome successfully with a reordering task. We have also proven that, with the use of morphological information, especially for a morphologically rich language like Kannada, the training data size can be reduced considerably with an improvement in performance.

3.2.1 Word Based Translation

As the name suggests, the words in an input sentence are translated word by word individually, and these words finally are arranged in a specific way to get the target sentence. The alignment between the words in the input and output sentences normally follows certain patterns in word based translation. This approach is the very first attempt in the statistical-based MT system that is comparatively simple and efficient. The main disadvantage of this system is the oversimplified word by word translation of sentences, which may reduce the performance of the translation system.

3.2.2 Phrase Based Translation

A more accurate SMT approach, called phrase-based translation (Koehn *et al*., 2003), was introduced, where each source and target sentence is divided into separate phrases instead of words before translation. The alignment between the phrases in the input and output sentences normally follows certain patterns, which is very similar to word based translation. Even though the phrase based models result in better performance than the word based translation, they did not improve the model of sentence order patterns. The alignment model is based on flat reordering patterns, and experiments show that this reordering technique may perform

well with local phrase orders but not as well with long sentences and complex orders.

3.2.3 Hierarchical Phrase Based model

By considering the drawback of previous two methods, Chiang (2005) developed a more sophisticated SMT approach, called the hierarchical phrase based model. The advantage of this approach is that hierarchical phrases have recursive structures instead of simple phrases. This higher level of abstraction approach further improved the accuracy of the SMT system.

3.3 Hybrid-based Translation

By taking the advantage of both statistical and rule-based translation methodologies, a new approach was developed, called hybrid-based approach, which has proven to have better efficiency in the area of MT systems. At present, several governmental and private based MT sectors use this hybrid-based approach to develop translation from source to target language, which is based on both rules and statistics. The hybrid approach can be used in a number of different ways. In some cases, translations are performed in the first stage using a rule-based approach followed by adjusting or correcting the output using statistical information. In the other way, rules are used to pre-process the input data as well as post-process the statistical output of a statistical-based translation system. This technique is better than the previous and has more power, flexibility, and control in translation.

Hybrid approaches integrating more than one MT paradigm are receiving increasing attention. The METIS-II MT system is an example of hybridization around the EBMT framework; it avoids the usual need for parallel corpora by using a bilingual dictionary (similar to that found in most RBMT systems) and a monolingual corpus in the TL (Dirix *et al*., 2005). An example of hybridization around the rule-based paradigm is given by Oepen. It integrates statistical methods within an RBMT system to choose the best translation from a set of competing hypotheses (translations) generated using rule-based methods (Oepen *et al*., 2007).

In SMT, Koehn and Hoang integrate additional annotations at the word-level into the translation models in order to better learn some aspects of the translation that are best explained on a morphological, syntactic, or semantic level (Koehn *et al*., 2007). Hybridization around the statistical approach to MT is provided by Groves and Way; they combine both corpus-based methods into a single MT system by incorporating phrases (sub-sentential chunks) from both EBMT and SMT into an SMT system (Groves *et al*., 2005). A different hybridization happens when an RBMT system and an SMT system are used in a cascade; Simard proposed an approach, analogous to that by Dugast, using an SMT system as an automatic post-editor of the translations produced by an RBMT system (Simard *et al*., 2007) (Dugast *et al*., 2007).

3.4 Example-based translation

The example-based translation approach is based on analogical reasoning between two translation examples, proposed by Makoto Nagao in 1984. At run time, an example-based translation is characterized by its use of a bilingual corpus as its main knowledge base. The example-based approach comes under the EMT system, which relies on large parallel aligned corpora.

Example-based translation is essentially translation by analogy. An EBMT system is given a set of sentences in the SL (from which one is translating) and their corresponding translations in the TL, and uses those examples to translate other, similar source-language sentences into the TL. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to be correct again. EBMT systems are attractive in that they require a minimum of prior knowledge; therefore, they are quickly adaptable to many language pairs.

A restricted form of example-based translation is available commercially, known as a translation memory. In a translation memory, as the user translates text, the translations are added to a database, and when the same sentence occurs again, the previous translation is inserted into the translated document. This saves the user the effort of re-translating that sentence, and is particularly effective when translating a new revision of a previously-translated document.

More advanced translation memory systems will also return close but inexact matches on the assumption that editing the translation of the close match will take less time than generating a translation from scratch. ALEPH, wEBMT, English to Turkish, English to Japanese, English to Sanskrit, and PanEBMT are some of the example-based MT systems.

3.5 Knowledge-Based MT

Knowledge-Based Machine Translation (KBMT) is characterized by a heavy emphasis on functionally complete understanding of the source text prior to the translation into the target text. KBMT does not require total understanding, but assumes that an interpretation engine can achieve successful translation into several languages. KBMT is implemented on the Interlingua architecture; it differs from other interlingual techniques by the depth with which it analyzes the SL and its reliance on explicit knowledge of the world.

KBMT must be supported by world knowledge and by linguistic semantic knowledge about meanings of words and their combinations. Thus, a specific language is needed to represent the meaning of languages. Once the SL is analyzed, it will run through the augmenter. It is the knowledgebase that converts the source representation into an appropriate target representation before synthesizing into the target sentence.

KBMT systems provide high quality translations. Nevertheless, they are quite expensive to produce due to the large amount of knowledge needed to accurately represent sentences in different languages. The English-Vietnamese MT system is one of the examples of KBMTS.

3.6 Principle-Based MT

Principle-Based Machine Translation (PBMT) Systems employ parsing methods based on the Principles & Parameters Theory of Chomsky‘s Generative Grammar. The parser generates a detailed syntactic structure that contains lexical, phrasal, grammatical, and thematic information. It also focuses on robustness, language-neutral representations, and deep linguistic analyses.

In the PBMT, the grammar is thought of as a set of language-independent, interactive well-formed principles and a set of language-dependent parameters. Thus, for a system that uses n languages, one must have n parameter modules and a principles module. Thus, it is well-suited for use with the interlingual architecture.

PBMT parsing methods differ from the rule-based approaches. Although efficient in many circumstances, they have the drawback of language-dependence and increase exponentially in rules if one is using a multilingual translation system. They provide broad coverage of many linguistic phenomena, but lack the deep knowledge about the translation domain that KBMT and EBMT systems employ. Another drawback of current PBMT systems is the lack of the most efficient method for applying the different principles. UNITRAN is one of the examples of PBMT.

3.7 Online Interactive Systems

In this interactive translation system, the user is allowed to suggest the correct translation to the translator online. This approach is very useful in a situation where the context of a word is unclear and there exists many possible meanings for a particular word. In such cases, the structural ambiguity can be solved with the interpretation of the user.

4. Major MT Developments in India: A Literature Survey

The first public Russian to English (Manning *et al.,* 2003) MT system was presented at Georgetown University in 1954 with a vocabulary size of around 250 words. Since then, many research projects have been devoted to MT. Nevertheless, as the complexity of the linguistic phenomena involved in the translation process together with the computational limitations of the time were made apparent, enthusiasm faded out quickly. Also, the results of two negative reports, namely ‘Bar-Hillel’ and ‘AL-PAC,’ had a dramatic impact on MT research in that decade.

During the 1970s, the focus of MT activity switched from the United States to Canada and Europe, especially due to the growing demands for translations within their multicultural societies. 'Mateo,' a fully-automatic system translating weather forecasts, enjoyed great success in Canada. Meanwhile, the European Commission installed a French-English MT system called 'Systran'. Other research projects, such as 'Eurotra,' 'Ariane,' and 'Susy,' broadened the scope of MT objectives and techniques. The rule-based approaches emerged as the correct path to successful MT quality. Throughout the 1980s, many different types of MT systems appeared with the most prevalent being those using an intermediate semantic language, such as the 'Interlingua' approach.

Lately, various researchers have shown better translation quality with the use of phrase translation. Most competitive SMT systems, such as CMU, IBM, ISI, and Google, use phrase-based systems with good results.

In the early 1990s, the progress made by the application of statistical methods to speech recognition, introduced by IBM researchers, was in purely-SMT models (Manning *et al*., 2003). The drastic increment in computational power and the increasing availability of written translated texts allowed the development of statistical and other corpus-based MT approaches. Many academic tools turned into useful commercial translation products, and several translation engines were quickly offered in the World Wide Web.

Today, there is a growing demand for high-quality automatic translation. Almost all of the research community has moved towards corpus-based techniques, which have systematically outperformed traditional knowledge-based techniques in most performance comparisons. Every year, more research groups embark on SMT experimentation, and there is regained optimism in regards to future progress within the community.

MT is an emerging research area in NLP for Indian languages, which started more than a decade ago. There have been number of attempts in MT for English to Indian languages and Indian languages to Indian languages using different approaches. The literature shows that the earliest published work was undertaken by Chakraborty in 1966 (Noone *et al*., 2003). Many government and private sector researchers, as well as individuals, are actively involved in the development of MT systems and have generated some reasonable MT systems. Some of these MT systems are in the advanced prototype or technology transfer stage, and the rest have been newly initiated. The main developments in Indian language MT systems are as follows.

4.1 ANGLABHARTI by Indian Institute of Technology, Kanpur (1991)

ANGLABHARTI is a multilingual machine aided translation project on translation from English to Indian languages, primarily Hindi, which is based on a pattern directed approach (Durgesh *et al*., 2000; Sinha *et al*., 1995; Ajai *et al*., 2009; Manning *et al*., 2003; Sudip *et al*.,

2005). The strategy in this MT system is better than the transfer approach and lies below the Interlingua approach. In the first stage, a pattern directed parsing is performed on the SL English, which generates a `pseudo-target' that is applicable to a set of Indian languages. Word sense ambiguity in the SL sentence also is resolved by a number of semantic tags. In order to transform the pseudo TL into the corresponding TL, the system uses a separate text generator module. After correcting all ill-formed target sentences, a post-editing package is used make the final corrections. Even though it is a general purpose system, it has been applied mainly in the domain of public health at present. The ANGLABHARTI system is currently implemented from English to Hindi translation called AnglaHindi which is web-enabled (http://anglahindi.iitk.ac.in) and has obtained good domain-specific results for health campaigns, successfully translating many pamphlets and medical booklets. At present, further research work is going on to extend this approach for English to Telugu/Tamil translation. The project is primarily based at IIT-Kanpur, in collaboration with ER&DCI, Noida, and has been funded by TDIL. Professor RMK Sinha, Indian Institute of Technology, Kanpur is leading this MT project.

4.2 ANGLABHARTI -II by Indian Institute of Technology, Kanpur (2004)

The disadvantages of the previous system are solved by introducing the ANGLABHARTI - II MT architecture system (Sinha *et al*., 2003). The different approach, a Generalized Example-Base (GEB) for hybridization in addition to a Raw Example-Base (REB), is used to improve the performance of the translation. Compared to the previous approach, this system first attempts a match in REB and GEB before invoking the rule-base at the time of actual usage. Automated pre-editing and paraphrasing steps are further improvements in the proposed new translation approach. The system is designed in a way that various submodules are pipelined in order to achieve more accuracy and robustness.

At present, the ANGLABHARTI technology has been transferred under the ANGLABHARTI Mission into eight different sectors across the country (Sudip *et al*., 2005). The main intention of this bifurcation is to develop Machine Aided Translation (MAT) systems for English to twelve Indian regional languages. These include MT from English to Marathi & Konkani (IIT, Mumbai): English to Asamiya and Manipuri (IIT, Guwahati): English to Bangla (CDAC, Kolkata): English to Urdu, Sindhi & Kashmiri (CDAC-GIST group, Pune): English to Malyalam (CDAC, Thiruvananthpuram): English to Punjabi (Thapar Institute of Engineering and Technology-TIET, Patiala) English to Sanskrit (Jawaharlal Nehru University - JNU, New Delhi): and English to Oriya (Utkal University, Bhuvaneshwar).

4.3 ANUBHARATI by Indian Institute of Technology, Kanpur (1995)

ANUBHARATI is a recently started MT system aimed at translating from Hindi to English (Durgesh *et al*., 2000; Sinha *et al*., 1995; Ajai *et al*., 2009; Sudip *et al*., 2005). Similar to the ANGLABHARTI MT system, ANUBHARATI is also based on machine aided translation in which a variation of the example-based approach, called a template or hybrid HEBM, is used. The literature shows that a prototype version of the MT system has been developed and the project is being extended for developing a complete system. The HEBMT approach takes advantage of pattern and example-based approaches by combining the essentials of these methods. One more added advantage of the ANUBHARATI system is that it provides a generic model for translation that is suitable for translation between any two Indian languages pair with a minor addition of modules.

4.4 ANUBHARATI-II by Indian Institute of Technology, Kanpur (2004)

ANUBHARATI-II is a revised version of the ANUBHARATI that overcomes most of the drawbacks of the earlier architecture with a varying degree of hybridization of different paradigms (Sudip *et al*., 2005). The main intention of this system is to develop Hindi to any other Indian languages, with a generalized hierarchical example-based approach. Nevertheless, while both ANGLABHARTI-I and ANUBHARTI-II did not produce the expected results, both systems have been implemented successfully with good results. Professor RMK Sinha, Indian Institute of Technology, Kanpur is leading this MT project.

4.5 Anusaaraka by Indian Institute of Technology, Kanpur and University of Hyderabad

To utilize the close similarity among Indian languages for MT, another translation system called Anusaaraka (Durgesh *et al*., 2000; Sudip *et al*., 2005), was introduced, which is based on the principles of Paninian Grammar (PG). Anusaaraka is a machine aided translation system that also is used on language access between these languages. At present, this system is applied to children's stories, and an Alpha version of the system has been developed already for language assessors from five regional languages Punjabi, Bengali, Telugu, Kannada, and Marathi into Hindi. The Anusaaraka MT approach mainly consists of two modules (Manning *et al*., 2003; Bharati *et al*., 1997). The first module is called Core Anusaaraka, which is based on language knowledge, and the second one is a domain specific module that is based on statistical knowledge, world knowledge, *etc*. That is, the idea behind Anusaaraka is different from other systems in that the total load is divided in-to parts. The machine carries out the language-based analysis of the text, and the remaining work, such as knowledge-based analysis or interpretation, is performed by the reader. The Anusaaraka project was funded by TDIL,started at IIT Kanpur, and later shifted mainly to the Centre for Applied Linguistics and

Translation Studies (CALTS), Department of Humanities and Social Sciences, University of Hyderabad. At present, the Language Technology Research Centre (LTRC) at IIIT Hyderabad is developing an English to Hindi MT system using the architecture of the Anusaaraka approach. This Anusaaraka project is being developed under the supervision of Prof. Rajeev Sangal and Prof. G U Rao.

4.6 Anusaaraka System from English to Hindi

The Anusaaraka system from English to Hindi preserves the basic principles of information preservation and load distribution of original Anusaaraka (Manning *et al*., 2003; Bharati *et al*., 1997). To analyze the source text, it uses a modified version of the XTAG based super tagger and light dependency analyzer that was developed at the University of Pennsylvania. The advantage of this system is that, after the completion of the source text analysis, the user may read the output and can always move to a simpler output if the system produces the wrong output or fails to produce output.

4.7 MaTra (2004)

MaTra is an English to Indian languages (at present Hindi) Human-Assisted translation system based on a transfer approach using a frame-like structured representation that resolves the ambiguities using rule-based and heuristics approaches (Durgesh *et al*., 2000; Sudip *et al*., 2005; Manning *et al*., 2003). MaTra is an innovative system, which provides an intuitive GUI, where the user visually can inspect the analysis of the system and can provide disambiguation information to produce a single correct translation. Even though the MaTra system is intended to be a general purpose system, it has been applied mainly in the domains of news, annual reports, and technical phrases. MaTra is an ongoing project and the system currently is able to translate domain-specific simple sentences. Current development is towards covering other types of sentences. The Natural Language group of the Knowledge Based Computer Systems (KBCS) division at the National Centre for Software Technology (NCST), Mumbai (currently CDAC, Mumbai) has undertaken the task developing the MaTra system and is funded by TDIL.

4.8 MANTRA by Centre for Development of Advanced Computing, Bangalore (1999)

The Mantra MT system is intended to perform translation for the domains of gazette notifications pertaining to government appointments and parliamentary proceeding summaries between English and Indian languages as well as from Indian languages to English, where source and TL grammars are represented using Lexicalized Tree Adjoining Grammar (LTAG) formalism (Durgesh *et al*., 2000; Sudip *et al*., 2005). The added advantage of this system is

that the system can also preserve the formatting of input Word documents across the translation. After the successful development of MANTRA-Rajyasabha, language pairs like Hindi-English and Hindi-Bengali translation already have started using the Mantra approach. The Mantra project is being developed under the supervision of Dr. Hemant Darbari and is funded by TDIL and the Department of Official Languages, Ministry of Home Affairs, Government of India.

4.9 UCSG-based English-Kannada MT by University of Hyderabad

Using the Universal Clause Structure Grammar (UCSG) formalism, the Computer and Information Sciences Department at the University of Hyderabad, under the supervision of Prof. K. Narayana Murthy, developed a domain-specific English-Kannada MT system (Durgesh *et al*., 2000; Sudip *et al*., 2005; Manning *et al*., 2003). This UCSG-based system is based on a transfer-based approach and has been applied to the translation of government circulars. The system work is done at the sentence level and requires post-editing. At its first step of translation, the source (English) sentence is analysed and parsed using UCSG parser (developed by Dr. K. Narayana Murthy). Then, using translation rules, an English-Kannada bilingual dictionary, and network based Kannada Morphological Generator (developed by Dr. K. Narayana Murthy), the system translates in-to the Kannada language. This project has been funded by government of Karnataka and work is going to improve the performance of the system. Later, the same approach was applied for English-Telugu translation.

4.10 UNL-based MT between English, Hindi and Marathi by Indian Institute of Technology, Mumbai

Universal Networking Language (UNL) MT between English, Hindi, and Marathi is based on the Interlingua approach (Durgesh *et al*., 2000; Sudip *et al*., 2005; Manning *et al*., 2003). Under the supervision of Prof. Pushpak Bhattacharya, IIT Bombay is the Indian participant in UNL, which is an international project of the United Nations University, aimed at developing an Interlingua for all major human languages in the world. In the UNL based MT, the knowledge of the SL is captured or converted into UNL form and reconverted from UNL to the TL, like Hindi and Marathi. The SL information is represented sentence by sentence which is later converted into a hypergraph having concepts as nodes and relations as directed arcs (Shachi *et al*., 2002). The document knowledge is expressed in three dimensions as word knowledge, conceptual knowledge, and attritute labels.

4.11 Tamil-Hindi Anusaaraka MT

The KB Chandrasekhar Research Centre of Anna University at Chennai is active in the area of Tamil NLP. A Tamil-Hindi language assessor has been built using the Anusaaraka formalism (Durgesh *et al*., 2000; Sudip *et al*., 2005; Manning *et al*., 2003). The group has developed a Tamil-Hindi machine aided translation system under the supervision of Prof. CN Krishnan, with a performance of 75%.

4.12 English-Tamil machine Aided Translation system

Recently, the NLP group also developed a prototype of English-Tamil Human Aided MT System (Manning *et al*., 2003; Dwivedi *et al*., 2010). The system mainly consists of three major components: an English morphological analyzer, a mapping unit, and the Tamil language morphological generator.

4.13 SHIVA MT System for English to Hindi

This project was developed jointly by the Indian Institute of Science, Bangalore, and International Institute of Information Technology, Hyderabad, in collaboration with Carnegie Mellon University based on an example-based approach (Sudip *et al*., 2005; Dwivedi *et al*., 2010). An experimental system has been released for experiments, trials, and user feedback and is publicly available.

4.14 SHAKTI MT System for English to Hindi, Marathi and Telugu

This is a recently started project that also was developed jointly by Indian Institute of Science, Bangalore, and International Institute of Information Technology, Hyderabad, in collaboration with Carnegie Mellon University (Sudip *et al*., 2005; Dwivedi *et al*., 2010). The system follows a hybrid approach by combining both rule and statistical-based approaches. An experimental system for English to Hindi, Marathi, and Telugu is publicly available for experiments, trials, and user feedback.

4.15 Anuvadak English-Hindi MT

Anuvadak 5.0 English to Hindi software is a general-purpose tool developed by the private sector company Super Infosoft Pvt Ltd., Delhi, under the supervision of Mrs. Anjali Rowchoudhury (Durgesh *et al*., 2000; Sudip *et al*., 2005; Manning *et al*., 2003; Dwivedi *et al*., 2010). The system has inbuilt dictionaries in specific domains and supports post-editing. If the corresponding target word is not present in the lexicon, the system has a facility to translate that source word into the target. The system can run in Windows and a demonstration version of the system is publicly available.

4.16 English-Hindi Statistical MT

A statistical-based English to Indian languages, mainly Hindi, MT system was started by IBM India Research Lab at New Delhi, using the same approach as its existing work on other languages (Durgesh *et al*., 2000; Manning *et al*., 2003).

4.17 English-Hindi MAT for news sentences

A rule-based English to Hindi Machine Aided Translation system was developed by Jadavpur University, Kolkata, under the supervision of Prof. Sivaji Bandyopadhyay (Durgesh *et al*., 2000). The system uses the transfer based approach and is currently working on domain specific MT system for news sentences.

4.18 A hybrid MT system for English to Bengali

Under the supervision of Prof. Sivaji Bandyopadhyay, a hybrid-based MT system for English to Bengali was developed at Jadavpur University, Kolkata, in 2004 (Dwivedi *et al*., 2010). The current version of the system works at the sentence level.

4.19 Hinglish MT system

In 2004, Prof. Sinha and Prof. Thakur developed a standard Hindi-English MT system called Hinglish by incorporating an additional level in the existing ANGLABHARTI-II and ANUBHARTI-II systems (Dwivedi *et al*., 2010). The system produced satisfactory results in more than 90% of the cases, except the case with polysemous verbs.

4.20 English to (Hindi, Kannada, Tamil) and Kannada to Tamil language-pair EBMT system (2006)

An example-based English to Hindi, Kannada, and Tamil, as well as Kannada to Tamil (Dwivedi *et al*., 2010), MT system was developed by Balajapally *et al*. (2006). A set of bilingual dictionaries comprised of a sentence dictionary, phrase-dictionary, word-dictionary, and phonetic-dictionary of parallel corpora of sentences, phrases, words, and phonetic mappings of words is used for the MT. A corpus size of 75,000 most commonly used English-{Hindi, Kannada and Tamil} sentence pairs are used for MT.

4.21 Punjabi to Hindi MT system (2007)

A direct word-to-word translation approach, a Punjabi to Hindi MT system, was developed by Josan and Lehal at Punjabi University, Patiala, and reported 92.8% accuracy (Dwivedi *et al*., 2010). In addition to the Punjabi-Hindi lexicon and morphological analysis, the system also consists of modules that support word sense disambiguation, transliteration, and post-processing.

4.22 MT System among Indian language - Sampark (2009)

Consortiums of institutions (including IIIT Hyderabad, University of Hyderabad, CDAC (Noida, Pune), Anna University, KBC, Chennai, IIT Kharagpur, IIT Kanpur, IISc Bangalore, IIIT Alahabad, Tamil University, Jadavpur University) started to develop MT systems among Indian languages, called Sampark and have already released experimental systems for {Punjabi, Urdu, Tamil, Marathi} to Hindi and Tamil-Hindi in 2009 (Dwivedi *et al*., 2010).

4.23 English to Bengali (ANUBAAD) and English to Hindi MT System by Jadavpur University

Using a phrasal example-based approach, Jadavpur University developed a domain- specific translation of English news to Bengali called ANUBAAD, with current system work at the sentence level (Sudip *et al*., 2005). Also, the university started to develop a translation system for English news headlines to Bengali using a semantics-example-based approach. Using the same architecture, the university also developed a MT system for English-Hindi, and the system works currently at the simple sentence level. Recently the university also started to develop an Indian languages (Bengali, Manipuri) to English MT system. These translation systems are developing under the supervision of Prof. Sivaji Bandyopadhyay. The university uses these translation systems for guiding students and researchers who work in the MT area.

4.24 Oriya MT System (OMTrans) by Utkal University, Vanivihar

Utkal University, Bhuvaneshwar is working on an English-Oriya MT system OMTrans under the supervision of Prof. Sanghamitra Mohanty (Sudip *et al*., 2005; Manning *et al*., 2003). In addition to the parser and Oriya Morphological Analyser (OMA), the system also consists of an N-gram based word sense disambiguation module.

4.25 English-Hindi EBMT system by IIT Delhi

The Department of Mathematics, IIT Delhi, under the supervision of Professor Niladri Chatterjee developed an example-based English-Hindi MT system (Sudip *et al*., 2005). They have developed divergence algorithms for identifying the divergence for English to Hindi example-based system and a systematic scheme for retrieval from the English-Hindi example base.

4.26 Machine Aided Translation by Centre for Development of Advanced Computing (CDAC), Noida

Using the Machine Aided Translation system approach, a domain-specific translation system for translating public health related sentences from English to Hindi was developed (Manning *et al*., 2003). The system supports the advantage of post-editing and reportes 60%

performance.

4.27 Hindi to Punjabi MT system (2009)

Goyal and Lehal of Punjabi University, Patiala, developed a Hindi to Punjabi MT system based on a direct word-to-word translation approach (Goyal *et al*., 2009; Dwivedi *et al*., 2010). The system consists of the following modules: pre-processing, a word-to-word Hindi-Punjabi lexicon, morphological analysis, word sense disambiguation, transliteration, and post-processing. They also have developed an evaluation approach for a Hindi to English translation system and have reported 95% accuracy. Still, work is being carried out to achieve a better system.

4.28 A Statistical MT Approach to Sinhala-Tamil Language (2011)

Ruvan Weerasinghe developed an SMT Approach to Sinhala-Tamil Language Translation (Weerasinghe *et al*., 2011). This work reports on SMT based translation performed between language pairs, such as the Sinhala-Tamil and English-Sinhala pairs. The experiments results show that current models perform significantly better for the Sinhala-Tamil pair than the English-Sinhala pair and prove that the SMT system works better for languages that are not too distantly related to each other.

4.29 An Interactive Approach for English-Tamil MT System on the Web (2002)

Dr. Vasu Renganathan, University of Pennsylvania, developed an interactive approach for an English-Tamil MT System on the Web (Samir *et al*., 2010). The system is set on a rule-based approach, containing around five thousand words in the lexicon and a number of transfer rules used for mapping English structures to Tamil structures. This is an interactive system in that users can update this system by adding more words into the lexicon and rules into the rule-base.

4.30 Translation system using pictorial knowledge representation (2010)

Samir Kr. Borgohain and Shivashankar B. Nair introduced a new MT approach for Pictorially Grounded Language (PGL) based on their pictorial knowledge (Samir *et al*., 2010). In this approach, symbols of both the source and the TLs are grounded on a common set of images and animations. PGL is a graphic language and acts as a conventional intermediate language representation. While preserving the inherent meanings of the SL, the translation mechanism can also be scalable into a larger set of languages. The translation system is implemented in such a way that images and objects are tagged with both the source and target language equivalents, which makes the reverse translation much easier.

4.31 Rule-based Reordering and Morphological Processing For English-Malayalam SMT (2009)

This is an attempt to develop a statistical-based MT for English to Malayalam language by a set of MTech students under the guidance of Dr. K P Soman (Rahul *et al*., 2009). In this approach, they showed that a SMT based system can be improved by incorporating the rule-based reordering and morphological information of source and target languages.

4.32 SMT using Joshua (2011)

A piloted SMT based English to Telugu MT (MT) System called "enTel" was developed by Anitha Nalluri and Vijayanand Kommaluri, based on Johns Hopkins University Open Source Architecture (JOSHUA) (Anitha *et al*., 2011). A Telugu parallel corpus from the Enabling Minority Language Engineering (EMILLE) developed by CIIL Mysore and English to Telugu Dictionary, developed by Charles Philip Brown, is considered for training the translation system.

4.33 Multilingual Book Reader

The NLP team, including Prashanth Balajapally, Phanindra Bandaru, Ganapathiraju, N. Balakrishnan and Raj Reddy, introduced a multilingual book reader interface for DLI that supports transliteration and good enough translation (Prashanth) based on transliteration, word to word translation and full-text translation for Indian language. This is a simple, inexpensive tool that exploits the similarity between Indian languages. This tool can be useful for beginners who can understand their mother tongue or other Indian languages, but cannot read the script, and for an average reader who has the domain expertise. This tool can be also be used for translating either the documents or the queries in a multilingual search purpose.

4.34 A Hybrid Approach to EBMT for English to Indian Languages (2007)

Vamshi Ambati and U Rohini proposed a hybrid approach to EBMT (EBMT) for English to Indian languages that makes use of SMT methods and minimal linguistic resources (Ambati *et al*., 2007). Currently work is going on to develop English to Hindi as well as other Indian language translation systems based on manual and a statistical dictionary built from an SMT tool using an example database consisting of source and target parallel sentences.

4.35 SMT by Incorporating Syntactic and Morphological Processing

Ananthakrishnan Ramanathan, Pushpak Bhattacharyya, Jayprasad Hegde, Ritesh M. Shah, and M. Sasikumar proposed a new idea to improve the performance of the SMT based MT by incorporating syntactic and morphological processing (Ananthakrishnan). In this contest, they proved that performance of a baseline phrase-based system can be substantially improved by i)

reordering the source (English) sentence as per target (Hindi) syntax, and (ii) using the suffixes of target (Hindi) words.

4.36 Prototype MT System from Text-To-Indian Sign Language (ISL)

This is a very different approach to MT that is intended for dissemination of information to the deaf people in India and was proposed by Tirthankar Dasgupta, Sandipan Dandpat, and Anupam Basu (Dasgupta *et al*. 2008; Harshawardhan *et al*., 2011). At present, a prototype version of English to Indian Sign Language has been developed and the ISL syntax is represented based on Lexical Functional Grammar (LFG) formalism.

4.37 An Adaptable Frame based system for Dravidian language Processing (1999)

In the proposed work, a different approach that makes use of the karaka relations for sentence comprehension is used in the frame-based translation system for Dravidian languages (Idicula *et al*., 1999). Two pattern-directed application-oriented experiments are conducted, and the same meaning representation technique is used in both cases. In the first experiment, translation is done from a free word order language to fixed word order one, where both the source and destination are natural languages. In the second experiment, however, the TL is an artificial language with a rigid syntax. Even though there is a difference in the generation of the target sentence, the results obtained in both experiments are encouraging.

4.38 English-Telugu T2T MT and Telugu-Tamil MT System (2004)

CALTS in collaboration with IIIT, Hyderabad; Telugu University, Hyderabad; and Osmania University, Hyderabad developed an English-Telugu and Telugu-Tamil MT system under the supervision of Prof. Rajeev Sangal (CALTS). The English-Telugu system uses an English-Telugu machine aided translation lexicon of size 42000 words and a wordform synthesizer for Telugu. The Telugu-Tamil MT system was developed based on the available resources at CALTS: Telugu Morphological analyzer, Tamil generator, verb sense disambiguator, and Telugu-Tamil machine aided translation dictionary. The performance of the systems is encouraging, and it handles source sentences of varying complexity.

4.39 Developing English-Urdu MT Via Hindi (2009)

R. Mahesh K. Sinha proposed a different strategy for deriving English to Urdu translation using an English to Hindi MT system (R. Mahesh *et al*., 2009). In the proposed method, an English-Hindi lexical database is used to collect all possible Hindi words and phrases. These words and phrases are further augmented by including their morphological variations and attaching all possible postpositions. Urdu is structurally very close to Hindi and this

augmented list is used to provide mapping from Hindi to Urdu. The advantage of this translation system is that the grammatical analysis of English provides all the necessary information needed for Hindi to Urdu mapping and no part of speech tagging, chunking, or parsing of Hindi has been used for translation.

4.40 Bengali-Assamese automatic MT system-VAASAANUBAADA (2002)

Kommaluri Vijayanand, S. Choudhury and Pranab Ratna proposed an automatic bilingual MT for Bengali to Assamese using an example-based approach (Kommaluri *et al.*, 2002). They used a manually created aligned bilingual corpus by feeding real examples using pseudo code. The quality of the translation was improved by preprocessing the longer input sentences and also via the backtracking techniques. Since the grammatical structure of Bengali and Assamese is very similar, lexical word groups are required.

4.41 Phrase based English-Tamil Translation System by Concept Labeling using Translation Memory (2011)

The Computational Engineering and Networking research centre of Amrita School of Engineering, Coimbatore, proposed an English-Tamil translation system. The system is set on a phrase-based approach by incorporating concept labeling using translation memory of parallel corpora (Harshawardhan *et al.*, 2011). The translation system consists of 50,000 English-Tamil parallel sentences, 5000 proverbs, and 1000 idioms and phrases, with a dictionary containing more than 2,00,000 technical words and 100,000 general words. The system has an accuracy of 70%.

4.42 Rule-based Sentence Simplification for English to Tamil MT System (2011)

This work is aimed at improving the translation quality of an MT system by simplifying the complex input sentences for an English to Tamil MT system (Poornima *et al.*, 2011). In order to simplify the complex sentences based on connectives, like relative pronouns or coordinating and subordinating conjunctions, a rule-based technique is proposed. In this approach, a complex sentence is expressed as a list of sub-sentences while the meaning remains unaltered. The simplification task can be used as a preprocessing tool for MT where the initial splitting is based on delimiters and the simplification is based on connectives.

4.43 Manipuri-English Bidirectional SMT Systems (2010)

Using morphology and dependency relations, a Manipuri to English bidirectional SMT system was developed by Thoudam Doren Singh and Sivaji Bandyopadhyay (Doren Singh *et al.*, 2010). The system uses a domain-specific parallel corpus of 10350 sentences from news for

training purposes and the system is tested with 500 sentences.

4.44 English to Kannada SMT System (2010)

P.J. Antony, P. Unnikrishnan and Dr. K.P Soman proposed an SMT system for English to Kannada by incorporating syntactic and morphological information (Unnikrishnan *et al*., 2010). In order to increase the performance of the translation system, we have introduced a new approach in creating the parallel corpus. The main ideas that we have implemented and proven effective in the English to Kannada SMT system are: (i) reordering the English source sentence according to Dravidian syntax, (ii) using the root suffix separation on both English and Dravidian words, and iii) use of morphological information that substantially reduces the corpus size required for training the system. The results show that significant improvements are possible by incorporating syntactic and morphological information into the corpus. From the experiments we have found that the proposed translation system successfully works for almost all simple sentences in their twelve tense forms and their negatives forms.

4.45 Anuvadaksh

This system is an effort of the English to Indian Language MT (EILMT) consortium. Anuvadaksh is a system that allows translating the text from English to six other Indian languages, *i.e*. Hindi, Urdu, Oriya, Bangla, Marathi, and Tamil. Anuvadaksh being a consortium based project has a hybrid approach that is designed to work with platform and technology independent modules.

This system has been developed to facilitate the multi-lingual community, initially in the domain-specific expressions of tourism, and it would subsequently foray into various other domains in a phase-wise manner. It integrates four MT Technologies:

Tree-Adjoining-Grammar (TAG) based MT.
SMT.
Analyze and Generate rules (Anlagen) based MT.
Example-based MT (EBMT).

4.46 Google Translate

Google Translate is a free translation service that provides instant translations between 57 different languages. Google Translate generates a translation by looking for patterns in hundreds of millions of documents to help decide on the best translation. By detecting patterns in documents that have already been translated by human translators, Google Translate makes guesses as to what an appropriate translation should be. This process of seeking patterns in large amounts of text is called "SMT".

4.47 English to Assamese MT System

An English to Assamese MT system is in progress (Sudhir *et al*., 2007). The following activities are in progress in this direction.

• The graphical user interface of the MT system has been re-designed. It now allows the display of Assamese text. Modifications have been made in the Java modules.

• The existing Susha encoding scheme has been used. In addition, a new Assamese font set has been created according to that of Susha font set. The system is now able to display properly consonants, vowels, and matras of Assamese characters properly.

• The mapping of the Assamese keyboard with that of Roman has been worked out.

• The process of entering Assamese words (equivalent of English words) in the lexical database (nouns and verbs) is in progress.

The system developed basically a rule-based approach and relies on a bilingual English to Assamese dictionary. The dictionary-supported generation of Assamese text from English text is a major stage in this MT. Each entry in the dictionary is supplied with inflectional information about the English lexeme and all of its Assamese equivalents. The dictionary is annotated for morphological, syntactic, and partially semantic information. It currently can handle translation of simple sentences from English to Assamese. The dictionary contains around 5000 root words. The system simply translates source language texts to the corresponding target language texts phrase to phrase by means of the bilingual dictionary lookup.

4.48 Tamil University MT System

Tamil University, Tanjore, initiated a machine oriented translation from Russian-Tamil during 1983-1984 under the leadership of Vice-Chancellor Dr. V.I Subramaniam (Sudhir *et al*., 2007). It was taken up as an experimental project to study and compare Tamil with Russian in order to translate Russian scientific text into Tamil. A team consisting of a linguist, a Russian language scholar, and a computer scientist were entrusted to work on this project. During the preliminary survey, both Russian SL and Tamil were compared thoroughly for their style, syntax, morphological level, *etc*.

4.49 Tamil-Malayalam MT System

Bharathidasan University, Tamilnadu, is working on translation between languages belonging to the same family, such as Tamil-Malayalam translation (Sudhir *et al*., 2007). The MT consists of the following modules that are in progress.

Lexical database- This will be a bilingual dictionary of root words. All the noun roots and verb roots are collected.

Suffix database- Inflectional suffixes, derivative suffixes, plural markers, tense markers, sariyai, case suffixes, relative participle markers, verbal participle markers, *etc* will be compiled.

Morphological Analyzer- This is designed to analyze the constituents of the words. It will help to segment the words into stems and inflectional markers.

Syntactic Analyzer- The syntactic analyzer will find the syntactic category, like Verbal Phrase, Noun Phrase, and Participle Phrase. This will analyze the sentences in the source text.

Table 1 below provides a summary of all 49 MT systems.

Table 1. Comparison of MT systems in India

Sr. No	MT System (Year)	Source-Target Language	Developer	Approach	Domain
1	ANGLABHARTI (1991)	English to Indian languages (primarily Hindi)	IIT, Kanpur	Pseudo-interlingua	General
2	ANGLABHARTI - II (2004)	English to Indian languages	IIT, Kanpur	Pseudo-interlingua	General
3	ANUBHARATI (1995)	Hindi to English	IIT, Kanpur	GEBMT	General
4	ANUBHARATI-II (2004)	Hindi to any other Indian languages	IIT, Kanpur	GEBMT	General
5	Anusaaraka (1995)	Punjabi, Bengali, Telugu, Kannada, and Marathi to Hindi.	IIT, Kanpur and University of Hyderabad	PG	General
6	Anusaaraka (1995)	from English to Hindi	IIT, Kanpur and University of Hyderabad	PG	General
7	MaTra (2004)	English to Indian languages (at present Hindi)	CDAC, Mumbai	Transfer based	General
8	MANTRA (1999)	English to Indian languages and Reverse	CDAC, Pune	TAG	Administration, office orders
9	UCSG-based MT	English-Kannada	University of Hyderabad	transfer based	government circulars
10	UNL-based (2003)	Between English, Hindi, and Marathi	IIT, Mumbai	Interlingua	General
11	Tamil-Hindi Anusaaraka MT	Tamil-Hindi	KBC Research Centre, Anna University,	PG	General
12	English-Tamil HAMT	English-Tamil	NLP group	HAMT	General

Sr. No	MT System (Year)	Source-Target Language	Developer	Approach	Domain
13	SHIVA (2004)	English to Hindi	IISc- Bangalore, IIIT Hyderabad, and Carnegie Mellon University	EBMT	General
14	SHAKTI (2004)	English to Hindi, Marathi and Telugu	IISc- Bangalore, IIIT Hyderabad, and Carnegie Mellon University	RBM	General
15	Anuvaadak	English-Hindi	Super Infosoft Pvt Ltd., Delhi	Not-Available	Not-Available
16	English-Hindi Statistical MT	English to Indian languages	IBM India Research Lab, New Delhi	EBMT & SMT	Not-Available
17	English-Hindi MAT	English to Hindi	Jadavpur University, Kolkata	transfer based	newssentences
18	Hybrid MT system	English to Bengali	Jadavpur University Kolkata	Hybrid	Sentence level
19	Hinglish MT system (2004)	Hindi - English	IIT-Kanpur	Pseudo interlingua	General
20	English to Indian and Kannada to Tamil language-pair EBMT system (2006)	i) English to Hindi, Kannada, and Tamil ii) Kannada to Tamil	Balajapally	Example-based	Most Commonly used sentences
21	Punjabi to Hindi MT system (2007)	Punjabi to Hindi	Punjabi University, Patiala	Direct word to word	General
22	Sampark (2007)9	Among Indian languages	Consortiums of institutions	CPG	Not-Available
23	ANUBAAD (2004)	English to Bengali and English to Hindi	Jadavpur University	RBMT & SMT	News Sentences
24	OMTrans	English-Oriya	Utkal University, Bhuvaneshwar	Not-Available	SchoolbookSentences
25	English-Hindi EBMT system	English-Hindi	IIT Delhi	Example-based, Divergence algorithms	Not-Available
26	Machine Aided Translation	English to Hindi	CDAC, Noida	Machine Aided Translation	Public health related sentences
27	Hindi to Punjabi MT system (2009)	Hindi to Punjabi	Punjabi University, Patiala	direct word-to-word	General

Sr. No	MT System (Year)	Source-Target Language	Developer	Approach	Domain
28	Sinhala-Tamil MT (2011)	Sinhala to Tamil	RuvanWeerasinghe	SMT based	General
29	English-Tamil MT on Web (2002)	English to Tamil	University of Pennsylvania	rule-based	General
30	Pictorial knowledge Based MT (2010)	English to Assamese	Samir Kr. Borgohain and Shivashankar B. Nair	pictorial knowledge	People not well versed in each other‟s languages
31	English-Malayalam Statistical MT (2009)	English to Malayalam	AMRITA University, Coimbatore	SMT based	General
32	enTel (2011)	English to Telugu	AnithaNalluri and VijayanandKommaluri	SMT based	Not-Available
33	Multilingual book reader interface for DLI	Translation for Indian languages	PrashanthBalajapally and Team	Word-to-Word Translation	documents or the queries
34	English to Indian Languages MT (2007)	English to Indian Languages	VamshiAmbati and Rohini U proposed ,	Example-based	Not-Available
35	Incorporating Syntactic and Morphological based MT	English-Hindi	AnanthakrishnanRamanathan and Team	Stasticalphrase-based	Not-Available
36	Text-To-Indian Sign Language (ISL) MT	English to Indian Language	TirthankarDasgupta, SandipanDandpat, and AnupamBasu	Lexical Functional Grammar (LFG) formalism	Deaf people in India
37	Dravidian language Processing System (199)	Dravidian language	SumamUMAM MARY IDICULA	Adaptable Frame based	Not-Available
38	English-Telugu T2T MT and Telugu-Tamil MT (2004)	English-Teluguand Telugu-Tamil	CALTS; IIIT-Hyderabad; Telugu University- Hyderabad, Osmania University-Hyderabad,	Not-Available	Not-Available
39	English-Urdu MT via Hindi (2009)	English-Urdu	R. Mahesh K. Sinha	Not-Available	Not-Available
40	VAASAANUBAADA (2002)	Bengali- Assamese	KommaluriVijayanand S Choudhury and PranabRatna	EBMT	News

Sr. No	MT System (Year)	Source-Target Language	Developer	Approach	Domain
41	Phrase based English - Tamil MT (2011)	English - Tamil	CEN, AMRITA University, Coimbatore	Phrase based	General
42	Sentence Simplification System for English to Tamil (2011)	English - Tamil	Not-Available	Rule-based	Not-Available
43	Manipuri-English Bidirectional MT (2010)	Manipuri-English and -English-Manipuri	ThoudamDoren Singh and SivajiBandyopadhyay	Statistical	news
44	English to Dravidian Language MT (2010)	English to Malayalam	CEN, AMRITA University, Coimbatore	SMT Based	Simple sentences
45	Anuvadaksh	English to six other Indian languages *i.e.* Hindi, Urdu, Oriya, Bangla, Marathi, Tamil	EILMT consortium	hybrid approach	Tourism
46	Google Translate	Translations between 57 different languages	Google	SMT	General
47	English to Assamese MT	English to Assamese	Not-Available	Rule-based	Not-Available
48	Russian-Tamil MT (1983-1984)	Russian-Tamil	Tamil University, Tanjore	Not-Available	scientific text
49	Tamil - Malayalam MT	Tamil - Malayalam	Bharathidasan University, Tamil Nadu	Not-Available	Not-Available

5. Conclusion

This survey described machine translation (MT) techniques in a longitudinal and latitudinal way with an emphasis on the MT development for Indian languages. Additionally, we tried to describe briefly the different existing approaches that have been used to develop MT systems. From the survey, we found that almost all existing Indian language MT projects are based on a statistical and hybrid approach. We also identified the following two reasons that most of the developed MT systems for Indian languages have followed the statistical and hybrid approach. The first reason is, since Indian languages are morphologically rich in features and agglutinative in nature, rule-based approaches have failed in many situations for developing full-fledged MT systems. Second the general benefits of statistical and hybrid approaches have encouraged researchers to choose these approaches to develop MT systems for Indian languages.

Reference

ALPAC. (1966). *Language and Machines: Computers in Translation and Linguistics*. A report by the Automatic Language Processing Advisory Committee (Tech. Rep. No. Publication 1416), 2101 Constitution Avenue, Washington D.C., 20418 USA: National Academy of Sciences, National Research Council.

Ambati, V., & Rohini, U. (2007). A Hybrid Approach to EBMT for Indian Languages. *ICON 2007*.

Badodekar, S. (2003). *Translation Resources, Services and Tools for Indian Languages*. Computer Science and Engineering Department, Indian Institute of Technology, Mumbai, 400019, India.

Balajapally, P., Bandaru, P., Ganapathiraju, M., Balakrishnan, N., & Reddy, R. (2006). Multilingual Book Reader: Transliteration, Word-to-Word Translation and Full-text Translation. In *VAVA 2006*.

Bharati, A., Chaitanya, V., Kulkarni, A. P., & Sangal, R. (1997). ANUSAARAKA: Machine Translation in Stages. *A Quarterly in Artificial Intelligence*, 10(3), 22-25.

Borgohain, S. K., & Nair, S. B. (2010). Towards a Pictorially Grounded Language for Machine-Aided Translation. *International Journal on Asian Language Processing*, 20 (3), 87-109.

CALTS in collaboration with, IIIT Hyderabad. English-Telugu T2T Machine Translation and Telugu-Tamil Machine translation System. Indo-German Workshop on Language technologies, AU-KBC Research Centre, Chennai, 2004 . www.au-kbc.org/dfki/igws/Machine_Translation.ppt.

Dasgupta, T., & Basu, A. (2008). An English to Indian Sign Language Machine Translation System, www.cse.iitd.ac.in/embedded/assistech/Proceedings/P17.pdf.

Dasgupta, T., Dandpat, S., & Basu, A. (2008). Prototype Machine Translation System From Text-To-Indian Sign Language. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 19-26.

Dave, S., Parikh, J., & Bhattacharya, P. (2001). Interlingua-based English-Hindi Machine Translation and Language Divergence. *Journal of Machine Translation*, 16(4), 251-304.

Dirix, P., Schuurman, I., & Vandeghinste V. (2005). Metis II: Example-based machine translation using monolingual corpora - system description. In P*roceedings of the 2nd Workshop on Example-Based Machine Translation*, 43-50.

Dugast, L., Senellart, J., & Koehn, P. (2007). Statistical post-editing on SYSTRAN's rule-based translation system. In *Proceedings of the Second Workshop on SMT*, 220-223.

Dwivedi, S. K., & Sukhadeve, P. P. (2010). Machine Translation System in Indian Perspectives. *Journal of Computer Science*, 6(10), 1111-1116.

Goyal, V., & Lehal, G. S. (2009). Evaluation of Hindi to Punjabi Machine Translation System. *IJCSI International Journal of Computer Science*, 4(1), 36-39.

Groves, D. & Way, A. (2005). Hybrid example-based SMT: the best of both worlds. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, 183-190.

Harshawardhan, R., Augustine, M. S., & Soman, K. P. (2011). Phrase based English - Tamil Translation System by Concept Labeling using Translation Memory. *International Journal of Computer Applications* (0975 - 8887), 20(3), 1-6.

Hutchins, J. (1993). The first MT patents. *MT News International*, 14-15.

Hutchins, J. (2005). The history of machine translation in a nutshell. http://www.hutchinsweb.me.uk/Nutshell-2005.pdf.

Hutchins, W. J., & Lovtskii, E. (2000). Petr Petrovich Troyanskii (1854-1950): A forgotten pioneer of mechanical translation. *Machine translation*, 15(3), 187-221.

IBM. (1954). *701 Translator.* IBM Archives online: Press release January 8th 1954, http://www-03.ibm.com/ibm/history/exhibits/701/701-translator.html.

Idicula, S. M. (1999). *Design and Development of an Adaptable Frame-based System for Dravidian Language.* Ph.D thesis, Department of Computer Science, COCHIN University of Science and Technology.

Jain, A. (2009). Machine Aided Translation Systems: *The Indian Scenario*. 2(6), 2009. www.iitk.ac.in/infocell/Archive/dirnov2/ techno_machine.html.

Koehn, P. & Hoang, H. (2007). Factored translation models. In Proceedings of the 2007 Joint Conference on Empirical Methods. In *NLP and Computational Natural Language Learning*, 868-876.

Mahesh, R., & Sinha, K. (2009). Developing English-Urdu Machine Translation Via Hindi. In *Third Workshop on Computational Approaches to Arabic Scriptbased Languages (CAASL3), MT Summit XII*, Ottawa, Canada.

Manning, C., & Schutze, H. (2003). Foundations of Statistical NLP. *Proceedings of HLT/NAACL*.

Mishra, S. K. (2007). *Sanskrit Karaka Analyzer for Machine Translation*. PhD. Thesis, Jawaharlal Nehru University.

Nalluri, A., & Kommaluri, V. (2011). SMT using Joshua: An approach to build 'enTel' system. *Language in India, Special Volume:Problems of Parsing in Indian Languages*, 11(5), 1-6. www.languageinindia.com.

Naskar, S., & Bandyopadhyay, S. (2005). Use of Machine Translation in India: Current Status. In *Proceedings of MT SUMMIT X*; September 13-15, 2005, Phuket, Thailand.

Noone, G. (2003). *Machine Translation - A Transfer Approach*, A project report, www.scss.tcd.ie/undergraduate/bacsll/bacsll_web/nooneg0203.pdf.

Oepen, S., Velldal, E., Lønning, J. T., Meurer, P., Rosen, V., & Flickinger, D. (2007). Towards hybrid quality-oriented machine translation on linguistics and probabilities in MT. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation*, 144-153.

Poornima, C., Dhanalakshmi, V., Kumar M. A., & Soman, K. P. (2011). Rule-based Sentence Simplification for English to Tamil Machine Translation System. *International Journal of Computer Applications* (0975 - 8887), 25(8), 38-42.

Rahul, C., Dinunath, K., Ravindran, R., & Soman, K. P. (2009). Rule-based Reordering and Morphological Processing For English-Malayalam SMT. *International Conference on Advances in Computing, Control, and Telecommunication Technologies*, 458-460.

Ramanathan, A., Bhattacharyya, P., Hegde, J., Shah, R. M., & Sasikumar, M. (2008). Simple Syntactic and Morphological Processing Can Help English-Hindi SMT. In *IJCNLP 2008*.

Rao, M. D. (2000). *Machine Translation in India: A Brief Survey*. www.elda.org/en/proj/scalla/SCALLA2001/SCALLA2001Rao.pdf.

Renganathan, V. (2002). An Interactive Approach to Development of English-Tamil Machine Translation System on the Web. *Tamil Internet 2002*, California, USA. 68-73. www.infitt.org/ti2002/hubs/ conference/papers.html.

Simard, M., Ueffing, N., Isabelle, P., & Kuhn, R. (2007). Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on SMT*, 203-206.

Singh, T. D., & Bandyopadhyay, S. (2010). Manipuri-English Bidirectional SMT Systems using Morphology and Dependency Relations. In *Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation*, 83-91, COLING 2010, Beijing.

Sinha, R. M. K. & Jain, A. (2003). AnglaHindi: An English to Hindi Machine-Aided Translation System. In *MT Summit IX*, New Orleans, Louisiana, USA, September, 2003.

Sinha, R. M. K., Sivaraman, K., Agrawal, A., Jain, R., Srivastava, R. & Jain, A. (1995). ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages. *IEEE International Conference on: Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century*, 1609-1614.

Unnikrishnan, P., Antony, P. J., & Soman, K. P. (2010). A Novel Approach for English to South Dravidian Language SMT System. *International Journal on Computer Science and Engineering (IJCSE)*, 02(08), 2749-2759.

Vijayanand, K., Choudhury, S., & Ratna, P. (2002). Vaasaanubaada Automatic Machine Translation Of Bilingual Bengali - Assamese News Texts. *Language Engineering Conference*, University of Hyderabad, India.

Weaver, W. (1999). Warren Weaver Memorandum, July 1949. *MT News International*, no. 22, July 1999, 5-6, 15.

Weerasinghe, R. (2011). A SMT Approach to Sinhala-Tamil Language Translation. citeseerx.ist.psu.edu /viewdoc/summary?doi= 10.1.1.78.7481, 2011.

Zhang, Y. (2006). Chinese-English SMT by Parsing. www.cl.cam.ac.uk/~yz360/ mscthesis.pdf.

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 1, March 2013, pp. 79-98

Emotion Co-referencing - Emotional Expression, Holder, and Topic

Dipankar Das*, and Sivaji Bandyopadhyay+

Abstract

The present approach aims to identify the *emotional expression*, *holder*, *topic,* and their co-reference from Bengali blog sentences. Two techniques are employed, one is a rule-based baseline system and the other is a supervised system that consists of different syntactic, semantic, rhetorical, and overlapping features. Different error cases have been resolved using rule-based post processing techniques. The evaluative vectors containing emotional expressions, *holders,* and *topics* are prepared from annotated blog posts as well as from system generated output. The evaluation metric, *Krippendorff's* α**,** achieves agreement scores of 0.53 and 0.67 for the baseline and supervised co-reference classification systems, respectively.

Keywords: Emotional Expression, Holder, Topic, Co-reference Agreement.

1. Introduction

In psychology and in common use, emotion is an aspect of a person's mental state of being, normally based on or tied to the person's internal (physical) and external (social) sensory feeling (Zhang *et al.*, 2008). Emotions, of course, are not linguistic features. Nevertheless, the most convenient access we have to them is through language (Strapparava & Valitutti, 2004). The identification of emotion expressed in the text with respect to the reader or writer is a challenging task (Yang *et al.*, 2009). A wide range of Natural Language Processing (NLP) tasks, from tracking users' emotion about products/events/politics as expressed in online forums or news to customer relationship management, use emotional information.

Currently, emails, blogs, chat rooms, online forums, and even Twitter are being considered as effective communication substrates to analyze the reaction of emotional

* Department of Computer Science & Engineering, National Institute of Technology (NIT), Meghalaya, Shillong 793003, India
E-mail: dipankar.dipnil2005@gmail.com
+ Department of Computer Science & Engineering, Jadavpur University, West Bengal, Kolkata 700032, India
E-mail: sivaji_cse_ju@yahoo.com

catalysts. A blog is a communicative and informative repository of text-based emotional content in the Web 2.0 (Yang *et al.*, 2007). In particular, blog posts contain instant views, updated views, or influenced views regarding single or multiple topics. Many blogs act as online diaries of the bloggers for reporting the blogger's daily activities and surroundings. Sometimes, the blog posts are annotated by other bloggers. In addition, a large collection of blog data is suitable for any machine learning framework.

It has been observed that three major components are crucial in determining the emotional slants from different perspectives: *Emotional Expression*, *Holder*, and *Topic*. Thus, the determination of the emotion *holder* and *topics* from the text helps us track and distinguish users' emotions separately on the same or different *topics*. *Emotional expression* (word or phrase) is the subjective counterpart that can be expressed by a directly affective word ("John is really *happy* enough") or using some indirect notion ("Dream of music is in their eyes and hearts"). The source or *holder* of an *emotional expression* is the speaker, the writer, or the experiencer (Wiebe *et al.*, 2005). Extraction of the emotion *holder* is important in discriminating between emotions that are viewed from different perspectives (Seki, 2007). By grouping opinion *holders* of different stances on diverse social and political issues, we can gain better understanding of the relationships among countries or among organizations (Kim & Hovy, 2006). *Topic*, however, is the real world object, event, or abstract entity that is the primary subject of the emotion or opinion intended by the *holder* (Stoyanov & Cardie, 2008a). *Topic* depends on the context in which its associated *emotional expression* occurs (Stoyanov & Cardie, 2008b). For example, the following Bengali sentence shows the *emotional expression*, its associated *holder*, and *topic*.

রাশেদ বলেছেন আপনার **কবিতাটা** পড়তে গিয়ে তার এই
(*Rashed*) (*bolechen*) (*apnar*) (*kobitata*) (*porte*) (*giye*) (*tar*) (*ei*)
সুন্দর **কৌতুকটা** মনে পড়ছিলো।
(*sundar*) (*koutukta*) (*mone*) (*porchilo*).

***Rashed** said that he was remembering this **beautiful comic** while reading your **poem**.*

Emotional Expression: সুন্দর কৌতুক (**beautiful comic**), *Holder*: < writer**,** রাশেদ (**Rashed**) >**,** *Topic*: কবিতা (**poem**).

In the above example, along with the *emotional expression* and *topic*, the *writer* of the blog post is also considered as a default *holder* according to our assumption, which is based on the nested source hypothesis (Wiebe *et al.*, 2005). Sometimes, the emotional sentences may or may not contain a direct clue for the *emotional expression*. There are certain example sentences that contain an *emotional expression* without a *holder* (*Tar Abhinoy ta satyoi khub*

akorshoniyo chilo [His acting was really attractive]). Nevertheless, the sentence contains a *topic* (*Tar Abhinoy* [His acting]). Sometimes, even the *emotional expressions* represent the potential *topics*. For example, the Bengali sentence, "*Ami Ramer **doohkho koshte** kende pheli.*" [*I fall into cry on the **sorrow** of Ram*] contains the text "*doohkho koshte*" [*sorrow*] that is treated as both the *emotional expression* and the *topic*. With the above examples and problems in mind, we hypothesize that the notion of user-topic co-references will facilitate both the manual and automatic identification of emotional views. Presently, we have assumed that the *holder* and *topic* are emotion co-referent if they share the same *emotional expression*.

The present task deals with the identification of users' emotions on different *topics* from an annotated Bengali blog corpus (Das & Bandyopadhyay, 2010a). Each sentence of the corpus is annotated with the emotional components, such as *emotional expression* (word/phrase), intensity (*high*, *general*, and *low*), associated *holder*, *topic*(s), and sentential tag of Ekman's six emotion classes (*anger, disgust, fear, happy*, *sad*, and *surprise*).

In this project, a simple rule-based baseline system is developed for identifying the *emotional expressions*, *holders*, and *topics*. The expressions are identified from shallow parsed sentences using Bengali WordNet Affect lists (Das & Bandyopadhyay, 2010b). A simple part-of-speech (POS) tag-based pattern matching technique is employed to identify the emotion *holders* and *topics* with respect to the *emotional expressions*. The presence of emotion *holders* and *topics* in the immediate neighborhood, shallow chunks that refer to their corresponding *emotional expressions*, gives the co-reference clues for the baseline system. The co-reference among the *emotional expressions*, *holders*, and *topics* is measured using Krippendorff's (2004) α metric. The error analysis suggests that the rich morphology and free phrase order nature of Bengali restricts the baseline system in capturing the *holder* and *topic* as well as disambiguating them in complex, compound, and passive sentences.

Thus, a Support Vector Machine (SVM) (Joachims, 1998) based supervised classifier is employed as well for co-reference identification. In this classifier, each of the input vectors containing *emotional expression*, associated *holder*, and *topic* is prepared from each of the annotated Bengali blog sentences. The feature vector is prepared based on the information present in the sentences containing lexical, syntactic, semantic, rhetorical, and overlapping features (word, part-of speech (POS), and Named Entity (NE)). Considering each of the input vectors as a unit to be coded in terms of the values of a variable, the standard Krippendorff's (2004) α metric produces a satisfactory score that outperforms the baseline system on the test set. This observation suggests that the adoption of error handling features, along with the features for syntax, semantics, and rhetorical structure, improves the performance of the co-reference identification reasonably. Different types of error cases have been analyzed, and we employed different rule-based post-processing techniques to solve the error cases. The rest of this paper is organized as follows. Section 2 describes the related work. The baseline

system is described in Section 3. The supervised framework with feature analysis is discussed in Section 4. Experiments and associated results are specified in Section 5. The error analysis and post processing techniques are discussed in Section 6. Finally, Section 7 concludes the paper.

2. Related Work

The current trend in the emotion analysis area is exploring machine learning techniques (Sebastiani, 2002; Mishne & Rijke, 2006) that consider the problem as text categorization or analogous to topic classification, which underscores the difference between machine learning methods and human-produced baseline models (Alm *et al.*, 2005). The affective text shared task on news headlines for emotion and valence level identification at SemEval 2007 has drawn focus to this field (Strapparava & Mihalcea, 2007). In order to estimate affects in text, the model proposed by Neviarouskaya *et al.* (2007) processes symbolic cues and employs natural language processing techniques.

Prior work in identification of opinion *holders* has sometimes identified only a single opinion per sentence (Bethard *et al.*, 2004) and sometimes several opinions (Choi, 2005). Identification of opinion *holders* for Question Answering with a supporting annotation task was attempted in Wiebe *et al.* (2005). Before that, another work on labeling the arguments of the verbs with their semantic roles using a novel frame matching technique was carried out in Swier and Stevenson (2004). Based on the traditional perspectives, another work discussed in Hu *et al.* (2006) uses an emotion knowledge base for extracting the emotion *holder*. The machine learning based classification task for "not *holder*," "weak *holder*," "medium *holder*," or "strong *holder*" is described in Evans (2007). Kim and Hovy (2006) identified the opinion holder with the topic from media text using semantic role labeling. An anaphor resolution based opinion *holder* identification method exploiting lexical and syntactic information from online news documents was attempted by Kim *et al.* (2007). The syntactic models of identifying the emotion *holder* for English emotional verbs were developed in Das and Bandyopadhyay (2010d).

In the related field of opinion *topic* extraction, different researchers have contributed their efforts (Kobayashi *et al.*, 2004; Nasukawa *et al.*, 2003; Popescu & Etzioni, 2005). Nevertheless, these works are based on lexicon look up and are applied to the domain of product reviews. The *topic* annotation task on the MPQA corpus is described in Stoyanov and Cardie (2008).

The method of identifying an opinion with its *holder* and topic from online news is described in Kim and Hovy (2006). The model extracts opinion *topics* associated with a specific argument position for subjective expressions signaled by verbs and adjectives. Similarly, the verb based argument extraction and associated *topic* identification is considered

in the present system. Nevertheless, opinion topic identification differs from topic segmentation (Choi, 2000). The opinion *topics* are not necessarily spatially coherent as there may be two opinions in the same sentence on different *topics*, as well as opinions on the same *topic* that are separated by opinions that do not share that *topic* (Stoyanov & Cardie, 2008). The authors established such a hypothesis by applying the technique of co-reference identification for topic annotation. In the case of our present system, the building of fine-grained *topic* knowledge based on the rhetorical structure and segmentation of *topics* using different types of lexical, syntactic, and overlapping features substantially reduces the problem of emotion *topic* distinction. It must be mentioned that the proposed method obtains a moderately more reliable alpha score in comparison to some related results in Stoyanov and Cardie (2008a).

Moreover, all of the aforementioned works have been attempted for English. Recent study shows that non-native English speakers support the growing use of the Internet[1]. In addition to that, a rapidly growing number of web users from multilingual communities have focused attention on improving multilingual search engines in respect to sentiment or emotion. This raises the demand for emotion analysis for languages other than English. Bengali is the sixth most popular language in the World[2], second in India, and the national language in Bangladesh, but it is less computerized than English. Works on emotion analysis in Bengali have started recently (Das & Bandyopadhyay, 2009a; 2010a). The comparative evaluation of the features on equivalent domains for Bengali and English language can be found in Das and Bandyopadhyay (2009b). To the best of our knowledge, at present, no such user-topic co-reference analysis of emotion has been conducted for Bengali or for other Indian languages. Thus, we believe that this work would meet the demands of user-*topic* focused emotion analysis systems.

3. Baseline System

A simple rule-based system has been designed to identify the *emotional expression*, *holder*, and *topic* from the sentences and their co-references. A simple neighboring chunk consideration approach that assumes that *emotional expression*, *holder*, and *topic* appear as neighboring chunks in a sentence has been introduced to identify the co-reference among the three components. The details of the system are as follows.

***Identifying Emotional Expression*:** The blog sentences are passed through an open sourced Bengali shallow parser[3]. This shallow parser gives different morphological

[1] http://www.internetworldstats.com/stats.htm

[2] http://www.ethnologue.com/ethno_docs/distribution.asp?by=size

[3] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

information (*root, lexical category of the root, gender, number, person, case, vibhakti, tam, suffixes, etc.*) that help in identifying the lexical patterns of the *emotional expressions*. The shallow parsed sentences are preprocessed to generate simplified lexical patterns (as shown below). We search through each of the component words from the chunks in the Bengali *WordNet Affect* lists (Das & Bandyopadhyay, 2010b). If any word present in a chunk is an emotion word (*e.g.* কৌতুক *koutuk* 'comic'), all of the words present in that extracted chunk are treated as the candidate seeds for an *emotional expression*. Identification of an *emotional expression* containing a single emotion word is straightforward. Nevertheless, we include all of the words of a chunk in order to identify long *emotional expressions*. Consecutive words that appear in a chunk and contain at least one emotion word also form an *emotional expression*. An example of a shallow parsed result follows.

((JJP **সুন্দর** **'sundar' [beautiful]** **JJ** <fs af='সুন্দর ,adj,,,,d,শূন্য,শূন্য'>)
(NP **কৌতুকটা** 'koutukta' [comic] **NN** <fs af=' কৌতুক ,v,,,,, টা , টা >))

In many cases, the components of a given *emotional expression* are separated by stop words (*e.g.* একটি *ekti* 'a,' ঐ *oi* 'that,' এই *ei* 'this'), conjunctions (*e.g.* এবং *ebong* 'and,' অথবা *athoba* 'or,' কিন্তু *kintu* 'but,' *etc.*), negations (*e.g.* নয় *noy* 'not,' না *na* 'neither,' *etc.*) or intensifiers (তাই *tai* 'so,' খুব *khub* 'very,' কম *kam* 'less,' বেশি *beshi* 'much'). Each *emotional expression* is tagged with any of Ekman's (1993) six emotions, based on the type of the component word of the *emotional expression* present in the Bengali *WordNet Affect* lists.

Identification of Emotion Holder: The baseline system considers the phrasal patterns containing similarity clues to identify the emotion *holders*. The patterns are grouped according to part-of-speech (POS) categories. It has been observed that the hints of grouping the patterns are present mostly in the user comment portions of the Bengali blog texts. The Bengali blog structure[4] has been designed well, and each of the user comment portions starts with a corresponding username. The username is the default hint that helps in capturing the first *holder* present in an anchoring vector representing the nested sources. In other cases, the POS tags of the shallow parsed sentences contain similar patterns at the lexical level. The Named Entities (NEs) that are tagged with NNPC (Compound proper noun), NNP (Proper noun), NNC (Compound common noun), NN (Common noun), or PRP (Pronoun) tags at the beginning of a sentence are tagged as the possible candidates for the emotion *holder*. The similarity pattern consists of two phrasal constituents, the subject and the verb. The portion of the sentence excluding the subject and the verb that contains the additional constituents of the

[4] www.amarblog.com/

sentence has been identified as the common portion (*Common_Portion*). As Bengali is a free phrase order language, the order between the verb and the common portion is not fixed.

A general POS level pattern, such as [<NNP/NNPC/NN/NNC/PRP> {<VBZ/VM><*Common_Portion*>}], is considered for capturing clues about an emotion *holder*. The components of the *Common_Portion* are assembled after the first occurring POS tags of types NNP, NNC, or PRP in the POS tagged sentence until the verb POS, like VBZ or VM, is reached. The remaining components present in the sentence after the verb are appended to the common portion (*Common_Portion*).

The similarity patterns mostly exist in simple sentences. Such information is difficult to obtain from complex or compound sentences under this scheme; thus, the system also fails to identify nested emotion *holder*s. A total of 59 complex sentences and 34 compound sentences were present in the test set of 503 sentences.

Identifying Emotion Topic: The shallow chunked texts formed by removing *emotional expressions* and *holders* were identified as the responsible spans that contain one or more potential emotion *topics*. Without attempting any typical strategy, the words containing only the POS tags of NNP or NNC of the shallow chunks were identified as the emotion *topics*.

Emotional Expression Holder / Topic Co-reference: The emotion *topic* is intended by the emotion *holder*, and the *topic* depends on the context in which its associated emotional expression occurs (Stoyanov & Cardie, 2008a). Based on this hypothesis, each identified *holder* and *topic* that is associated with an *emotional expression* in a sentence is termed as co-referent if it shares the same *emotional expression* with the others. A rule-based technique has been adopted to identify the co-reference between the *holder* and *topic* with respect to an *emotional expression* if the chunks that are responsible for emotion *holder* or *topic* are the immediately neighboring chunks of that *emotional expression*.

Evaluation: The three identified components were stored in a vector. The evaluation of the vectors was carried out using Krippendorff's (2004) α metric by considering each of the vectors as the unit to be analyzed. Two vectors were filled up by the emotional components that were acquired from the annotated sentences and system generated results. Considering the annotated and system generated outputs as two separate raters, we used the number of identified components as the values to be assigned for each vector. We evaluated the system through the help of inter-rater agreement and measured the performance of the system using Krippendorff's (2004) α metric. The inter-rater agreement produces an α score of 0.53 on the test set. This is a standard metric employed for inter-annotator reliability studies. Krippendorff's α is a theoretically founded measure with a nice probabilistic interpretation. It was designed to measure the reliability of coding agreement, and the generalization of this metric was used as the evaluation metric for identifying co-reference in opinion *topic*

annotation (Stoyanov & Cardie, 2008a). Krippendorff's alpha is applicable to any number of coders (each assigning one value to one unit of analysis); to incomplete (missing) data; to any number of values available for coding a variable; and to binary, nominal, ordinal, interval, ratio, polar, and circular metrics (Levels of Measurement). In addition, it adjusts itself to small sample sizes of reliability data. We have concentrated only on nominal alpha as we have considered the strings of names.

It is observed that some sentences may or may not contain all three emotional components. Hence, three out of four values for each of the raters are assigned based on the number of annotated or acquired emotional components from the gold standard and system-generated data, respectively. One value has been considered for undetermined cases. If no annotated or system generated emotional component is tagged or acquired, the corresponding vector unit is considered as incomplete or as containing missing data. The metric, nominal alpha, produces an α score of 0.53 for measuring the agreement between the annotated and system generated data. The lower score of α, along with the availability of different features in the corpus, also motivated us to adopt a machine learning framework.

4. Supervised Framework

The *Topic* co-reference resolution resembles another well-known problem in NLP - the noun phrase (NP) co-reference resolution that considers machine learning frameworks (Soon *et al.*, 2001; Ng & Cardie, 2002). Therefore, we adopted a Support Vector Machine (SVM) (Joachims, 1998) based standard machine learning approach for identifying *holder-topic* co-reference from the perspective of *emotional expressions*, where the input vectors contain emotional expressions, *holders*, and *topics*. The training and classification processes for SVM were carried out by YamCha toolkit[5] and TinySVM-0.07[6], respectively. The system was trained with 2234 sentences. The best feature set was identified using 630 development sentences. An Information Gain Based Pruning (IGBP) was applied to the development set, and it improved the performance of the supervised system significantly.

Feature plays a crucial rule in the SVM framework. By manually reviewing the blog data and different language specific characteristics, word level and context level features have been selected heuristically for our classification task. The heart of our method is to give an input vector containing the *emotional expression*, *holder*, and *topic*, and the goal of the classifier is to determine whether the co-reference exists among the available components of the vector or not. Therefore, we have considered five different classes for identifying the co-reference between any pair of components. The classes are Expression-*Holder* (EH), Expression-*Topic* (ET),

[5] http://chasen-org/~taku/software/yamcha/

[6] http://chasen.org/~taku/software/TinySVM/

Holder-*Topic* (HT), Expression-*Holder*-*Topic* (EHT), and *none*. We use the manually annotated corpus (Das & Bandyopadhyay, 2010a) to train the classifier automatically. We construct each training example for each input vector. The co-reference identification relies on the expressiveness of the features used to describe the training example. We use the following four categories of features: lexical, syntactic, semantic, rhetorical, and overlapping features.

4.1 Lexical Features

Parts-of-Speech **(POS)**: We are interested in the *noun*, *adjective*, *verb*, and *adverb* words as these are emotion informative constituents. The POS features are extracted from the shallow parsed results used by the baseline system.

Negations **(NEG)**: Negative words that are annotated in the corpus (Das & Bandyopadhyay, 2010a) (নয় *noy* 'not,' না *na* 'neither,' *etc*.) were considered as a separate feature.

Conjunctions **(CONJ)**: The *Conjunctions* were annotated in the emotion corpus (Das & Bandyopadhyay, 2010a). The conjunctions were used as features (*e.g*. এবং *ebong* 'and,' অথবা *athoba* 'or,' কিন্তু *kintu* 'but,' *etc*.) for training and testing.

Punctuation Symbols **(Sym)**: Symbols, such as (,), (!), (?), are often used in single or multiple numbers to emphasize *emotional expressions* and are considered crucial clues for identifying emotional presence in a sentence. Thus, a special feature for such symbols was added in the active feature set for training and testing of the supervised system.

Emoticons **(emot_icon)**: Emoticons (☺,☹,😐) and their consecutive occurrence generally contribute real sentiment to the *emotional expressions* that precede or follow them. Like punctuation symbols, emoticons were also included as a feature. A knowledge base for emoticons has been prepared by experts after minutely analyzing the Bengali blog data. Each image link of an emoticon in the raw corpus was mapped to its corresponding textual entity in the tagged corpus with its proper emotion types using the knowledge base (Das & Bandyopadhyay, 2009).

4.2 Extraction of Subcategorization Frames for Identifying Syntactic Features

We augment the knowledge of subcategorization frames or syntactic frames for identifying emotion *holders* and *topics.* The identification of syntactic frames is not straightforward. The detailed methodology is as follows.

Verb Identification: The words tagged as main verb (VM) and belonging to the verb group chunk (VGNF) in the corpus are identified (*e.g*. **ভালোবাসা** *bhalobasa* 'love') as simple verbs from the shallow parsed sentences. In cases of compound or conjunct verbs, patterns like {[XXX] (NN) [YYY] (VM)} are retrieved (*e.g*. VGNF {[**আনন্দ** *ananda*] (NN) [**করা** *kara*] (VM)}

means *enjoy*). The light verbs [YYY] tagged with 'VM' generally occur in any inflected form. Different suffixes may be attached to a simple verb or light verb depending on various features, like tense, aspect, and person. An in-house Bengali stemmer with an accuracy of 97.09% used a suffix list to identify the stem forms of the simple and light verbs.

English Equivalent Synset Identification: The determination of an equivalent English synset of a Bengali verb was carried out using a Bengali to English bilingual dictionary[7]. The method to extract the English equivalent synsets of the Bengali verbs was based on the work done in Banerjee *et al.* (2009). We have identified the English equivalent verb synsets of the Bengali verb entries that are present in the dictionary. For example, the dictionary entries for the conjunct verb **আনন্দ করা** *ananda kara* 'enjoy' are as follows.

< আনন্দ করা **v**. to ***rejoice***; to ***make merry***….>

Different synonyms for a Bengali verb having the same sense are separated using "," and different senses are separated using ";" in the dictionary. The synonyms, including similar senses of the target verb, were extracted from the dictionary and yielded a set called the English Equivalent Synset (EES). In the above example, two English Equivalent Synsets (EES) are extracted for the conjunct verb আনন্দ *ananda* করা *kara* 'enjoy'.

English Equivalent Frame Identification: It also has been found that each of the English Equivalent Synsets (EES) occurs in each separate class of English VerbNet (Kipper-Schuler, 2005). VerbNet associates the semantics of a verb with its syntactic frames and combines traditional lexical semantic information, such as *thematic roles* and *semantic predicates*, with *selectional restrictions*. Member verbs in the same VerbNet class share common syntactic frames; thus, they are believed to have the same syntactic behavior. The VerbNet files containing member verbs and possible subcategorization frames are stored in XML file format. Hence, the XML files of VerbNet were pre-processed to build up a general list that contains all verbs, their classes, and possible subcategorization frames (primary as well as secondary). This preprocessed list was searched to extract the present subcategorization frames for each verb (*e.g*. *love*) of the English Equivalent Synsets (EES) (*e.g*. love) corresponding to the Bengali verb. These extracted subcategorization frames are believed to be the valid set of argument structures for the Bengali verbs (Banerjee *et al.*, 2010).

Frame Matching: On the other hand, the shallow parsed Bengali sentences are passed through a rule based *phrasal-head* extraction module to identify the phrase level argument structures of the sentences corresponding to the position of the verbs. The extracted *head part* of every

[7] http://home.uchicago.edu/~cbs2/banglainstruction.html

phrase from a parsed sentence is considered as a component of its sentential argument structure. If an acquired argument structure for a Bengali emotional sentence is matched with any of the available extracted frames of English VerbNet, the *thematic role* based *holder* (*Experiencer, Agent, Actor, Beneficiary*, *etc*.) and *topic* (*Topic*, *Theme*, *Event, etc*.) information associated with the English frame syntax is mapped to the appropriate slot of the acquired Bengali argument structure. Tag conversion routines were developed to transform the POS of the system generated argument structures into the POS of the VerbNet frames. The phrase level similarity between these two languages helps in identifying the subcategorization frames (Banerjee *et al.*, 2009). An example follows:

রাশেদ অনুভব করেছিল যে রামের সুখ অন্তহীন

(*Rashed*) (*anubhob*) (*korechilo*) (*je*) (*Ramer*) (*sukh*) (*antohin*)

Rashed felt that Ram's pleasure is endless.

Vector: < EH_রাশেদ, EH_*রাম*, ET_সুখ >

Acquired Argument Structure: [NNP VM DET-*je* S]

The argument structure contains a sentential complement "S" started by যে –*je* with DET type POS. The argument structure is acquired for the Bengali conjunct verb অনুভব করা *anubhab kara* 'feel'. One of the extracted VerbNet frame syntax containing –*that* type sentential complement for the equivalent English verb *feel* is as [<NP value="*Experiencer*" > </VERB> < S-*that* (Sentential –*that* Complement)>]. As the acquired argument structure matches the extracted VerbNet frame syntax, the *holder* related roles (*e.g*. *Experiencer*) associated with the VerbNet frame was mapped to the equivalent phrase in the acquired argument structure of the Bengali sentence. The phrase (**রাশেদ**) is now considered as a candidate of emotion *holder*. Additionally, the sentential complement portion is also passed through the syntactic model for obtaining any implicit emotion *holders*. The case markers in Bengali are required to identify the emotion *holder*s as the case markers give the useful hints to capture the *selectional restrictions* that play a key role in distinguishing the emotion *holder*s from other valid alternatives.

4.3 Semantic Features

Emotion/Affect Words **(EW)**: The presence of a word in the Bengali *WordNet Affect* lists (Das & Bandyopadhyay, 2010b) identifies the emotion/affect words. The tagged affect words are considered as both lexical and semantic features in the case of handling the *emotional expressions*.

Intensifiers **(INTF)**: The Bengali *SentiWordNet* was developed by replacing each word entry in the synonymous set of the English *SentiWordNet* (Esuli & Sebastiani, 2006) by its possible

Bengali synsets using the English to Bengali bilingual dictionary that was developed as part of the EILMT project[8]. The chunks containing JJ (adjective) and RB (adverb) tagged elements were considered to be intensifiers. If the intensifier was found in the *SentiWordNet*, then the positive and negative scores of the intensifier were retrieved from the *SentiWordNet*. The intensifier is classified into the list of positive (pos) (**INTF*pos***) or negative (neg) (**INTF*neg***), for which the average retrieved score is higher. The intensifiers play an important role in identifying the lexical association among the component words of an *emotional expression* and linking the emotion components based on their POS.

Multiword Expressions: *Reduplicated* words (সন্দ সন্দ *sanda sanda* [doubt with fear]) and *Idioms* (তাসের ঘর *taser ghar* [weakly built], গৃহদাহ *grrihadaho* [family disturbance]), which were annotated in the Bengali emotion blog corpus (Das & Bandyopadhyay, 2010a), have been considered as semantic features for the *emotional expressions*.

4.4 Rhetoric Features

The present task acquires the rhetorical components, such as *locus*, *nucleus*, and *satellite* (Mann & Thompson, 1988), from a sentence, as these rhetorical clues help in identifying the individual *topic* spans. The part of the text span containing an annotated *emotional expression* is considered as *locus*. Primarily, the separation of *nucleus* from *satellite* is done based on the punctuation marks (,), (!), (?). Frequently used *discourse markers* (যেহেতু *jehetu* 'as,' যেমন *jemon* '*e.g.*,' কারণ *karon* 'because,' মানে *mane* 'means') and *causal verbs* (ঘটায় *ghotay* 'caused') also act as useful clues if they are explicitly specified in the text. Stoyanov and Cardie (2008a) mentioned that the *topic* depends on the context in which its associated *emotional expression* occurs. If any word of an *emotional expression* co-occurs with any word element of the *nucleus* or *satellite* in the same shallow chunk, the feature is considered a *common rhetoric similarity*. Otherwise, the feature is considered a *distinctive rhetoric similarity*. The features aim to separate emotion *topics* from non-emotion *topics* as well as the individual *topic* from an overlapped *topic* region.

4.5 Overlapping Features

Word Overlap: This feature is *true* if any two *topic* spans contain a common word.

Part-of-Speech Overlap: The verb, noun, adjective, and adverb are considered as overlapping informative constituents.

NP Co-reference: This binary feature is *True* if the two chunks contain NPs that are determined to be co-referent by applying a rule of *common rhetoric similarity*.

[8] English to Indian Languages Machine Translation (EILMT) is a TDIL project undertaken by the consortium of different premier institutes and sponsored by MCIT, Govt. of India.

Named Entity (NE): Each of the sentences is passed through a Named Entity Recognizer (Ekbal & Bandyopadhyay, 2008) to identify the named entities in that sentence.

If any word is tagged as a named entity (NE), a feature is assigned for either emotion *holder* or *topic*. If, however, the word is present in *satellite* and not tagged as an emotion *holder* (*EH*) feature, the word is selected as a potential candidate for *topic*. This distinguishing feature is considered for identifying the *holder* and *topic* separately from an NE overlapped context.

5. Experimental Results

The combination of multiple features in comparison with a single feature generally shows a reasonable performance enhancement of any classification system. The impact of different features and their combinations was measured on the development set of 630 sentences. Different unigram and bi-gram context features (word and POS tag level) and their combinations were generated from the training corpus as well. We added each feature into the active feature list one at a time to see if the inclusion of a feature in the existing feature set improved the *F-Score* of the system on the development set. The final active feature set was applied to the test data. During the SVM-based training phase, the current token word with the three previous and three following words and their corresponding POS, along with negation or intensifier, were selected as context features for that word. We used Krippendorff's (2004) alpha (as discussed in Section 3) for measuring the performance of the system. The importance of incorporating the features was examined through Information Gain (*InfoGain*). All of the results were obtained by the 10 fold cross validation method.

Information Gain Based Pruning (IGBP): This decision technique was used to measure the importance of a feature (X) with respect to the class attribute (Y). Formally, the information gain of a feature X with respect to a class attribute Y is the reduction in uncertainty about the value of Y when we know the value of X:

$$InfoGain\ (Y;X) = entropy(Y) - entropy(Y|X)$$

where X and Y are discrete variables taking values $\{x_1, x_2,....,x_m\}$ and $\{y_1,y_2,....,y_n\}$, respectively. The *Entropy*(Y) is defined as:

$$Entropy\ (Y) = - \sum_{i=1 \text{ to } n} P(Y=_{yi})\ log_2\ P(Y=_{yi})$$

The conditional entropy of Y given X is defined as:

$$Entropy\ (Y|X) = - \sum_{i=1 \text{ to } m} P(X=_{xj})\ Entropy\ (Y|X=_{xj})$$

The features with high Information Gain (*InfoGain*) reduce the uncertainty about a class to the minimum. In our experiment on the development set, all of the features except the features for the *causal verbs* and *distinctive rhetoric similarity* achieved a high Information Gain (*InfoGain*). The word features (*e.g.* non-emotional words, such as *gather*, *seem*, *etc.*) were not

considered based on a threshold of 0.5.

The metric, nominal alpha produced an α score of 0.53 between the annotated and system generated data. Generally, the alpha α score aims to probabilistically capture the agreement of annotated data and separate it from the chance of agreement. The baseline score achieved for the overall agreement was 0.53, which is below the generally accepted level, while α for the supervised system was 0.63, which is moderately acceptable and reliable. The scores of α for the baseline system and supervised system, along with some important features, their combinations, and pruning steps, are shown in Table 1. The α score loses its probabilistic interpretation due to the way it is adapted to the problem of co-reference classification. It is observed that the score of α increased rapidly while considering the syntactic, rhetorical, and overlapping features. The overlapping features also cause problems because of the free phrase order characteristics of the Bengali language. The overlapping context of *emotional expression* and *topic* generates errors. Nevertheless, the application of Named Entities (NEs) reduces the problem of distinguishing *holder* and *topic*.

Table 1. Krippendorfff's α for different feature combinations and pruning.

Features	Krippendorff's α
Baseline System	**0.5344**
Supervised System (Lexical Features)	0.3561
Supervised System (Syntactic Features)	0.4002
Supervised System (Semantic Features)	0.3215
Supervised System (Rhetorical Features)	0.4176
Supervised System (Overlapping Features)	0.2345
Supervised System (Lexical+Syntactic)	0.4890
Supervised System (Syntactic+Rhetoric)	0.5012
Supervised System (Syntactic+Semantic+Rhetoric)	0.5201
Supervised System (Lexical+Syntactic+Rhetoric)	0.5421
Supervised System (Lexical+Syntactic+Semantic+Rhetoric+Overlapping)	**0.6121**
Supervised System (All Features) + IGBP	**0.6332**

6. Error Analysis

The error analysis was conducted on the development set of 630 sentences. We incorporated different rule-based post processing techniques for handling the error cases, and the system achieved an alpha score of 0.67. Four types of error cases were identified, and four different rules were proposed to reduce the error cases.

Case 1: *Appositive Use:* The implicit emotion *holders* may be present in a sentence. (*e.g.* রাম 'Ram' in the case of রামের সুখ 'Ram's pleasure'). The identification of the emotion *holder* at the sentence level requires the knowledge of two basic constraints (*implicit* and *explicit*) separately. The *explicit* constraints identify the single prominent emotion *holder* that is directly involved with the *emotional expression*, whereas the *implicit* constraints identify all direct and indirect nested sources as emotion *holder*s. The following example contains the emotion *holder* নাসরিন সুলতানা (*Nasreen Sultana*) based on *implicit* constraints.

Holder: < গেদু চাচা, নাসরিন সুলতানা >

গেদু চাচা বলে, না গো বোন , আমি **নাসরিন সুলতানা**র

(**Gedu ChaCha**) (bole) : (na) (go) (bon) , (ami) (**Nasreen Sultanar**)

দুঃখের কথাতে *কেঁদে* ফেলি।

(dookher) (kathate) (*kende*) (feli)

Gedu Chacha *says, no my sister, I fall into cry on the sad speech of* ***Nasreen Sultana****.*

Solution: We considered the suffixes that are determined from the shallow parsed phrases to identify the appositive cases. In the above example, the appositive case (*e.g.* রামের সুখ (*Ram's pleasure*)) is also identified and placed in the vector by removing the inflectional suffix (-এর *-er* in this case). Sometimes, the vibhakti and tam information also play effective roles in identifying emotion holders.

Case 2: *Anaphoric Presence of Holders*: Another similar problem is identified in the above example. The emotion *holders* are sometimes referred to via anaphors. Sometimes, the candidate anaphors are linked with the *emotional expressions* instead of the actual emotion *holders*. The actual emotion *holder* গেদু চাচা 'Gedu ChaCha' expresses the emotion in a clause that is represented by the anaphor আমি ami 'I' in another clause.

Solution: The sentences of user comments in the adopted blog corpus contain a special default phrasal pattern that helps in identifying the emotion holders ([<Named Entity> <say>] *e.g.* গেদু চাচা বলে: (*Gedu ChaCha bole*), রাশেদ বলেছেন (*Rashed bolechen*)**,** and সায়ন বলেছে (*Sayan bolechhe*)). Hence, if a pronoun is present with an *emotional expression*, the preceding Named Entities of such a default phrasal pattern are considered as the emotion *holders*.

Case 3: *Multiple Holders and Topics:* The complex or compound sentences contain more than one clause, and each of the clauses may contain individual *emotional expressions*. The *holders* and *topics* associated with the *emotional expressions* in all of the clauses need fine-grained study of the sentential structures. The following example shows that two *emotional expressions* (দুঃখ *dookkha* 'sorrow' and আনন্দ *ananda* 'happy') contain two different

holders (গেদু চাচা *Gedu ChaCha* and চাচি *Chachi*).

গেদু চাচার ***দুঃখ*** থাকা সত্ত্বেও **চাচি** ***আনন্দ*** করে সবাইকে
(**Gedu ChaChar**) (***dookkha***) (thaka) (satweo) (**Chachi**) (***ananda***) (kare) (sabaike)
নিয়ে থাকে ।
(niye) (thake)
*Though **Gedu Chacha** has sorrow, **Chachi** lives happily with all.*

Solution: As the complex or compound sentences contain more than one clause and each of the clauses contains individual *emotional expressions*, we consider the sentential rhetorical structure. Instead of identifying rhetorical relations (Mann & Thompson, 1988), the present task acquires the rhetorical components, such as *locus*, *nucleus*, and *satellite* from a sentence, as these rhetoric clues help in identifying the individual *holder* and *topics* associated in each clause of the sentence. The part of the text span containing the *emotional expression* is considered as *locus*. Primarily, the separation of *nucleus* from *satellite* is done based on the punctuation marks (,), (!), (?). Frequently used *discourse markers* (যেহেতু *jehetu* 'as,' যেমন *jemon* '*e.g.*,' কারণ *karon* 'because,' মানে *mane* 'means') and *causal verbs* (ঘটায় *ghotay* 'caused') also are useful clues if they are explicitly specified in the text and present in a manually prepared seed list. If any word in the *emotional expression* co-occurs with any word element of the *nucleus* or *satellite* in the same chunk, the feature is considered a *common rhetoric similarity.* Otherwise, the feature is considered a *distinctive rhetoric similarity*. The chunks identified by the syntactic system as the holder and topic and tagged as *common rhetoric similarity* are only considered for each of the clauses of a sentence. For this reason, all possible holders and topics associated to all of the clauses of a sentence are identified by the syntactic system.

Case 4: *Overlapping Topic Spans:* It is observed that the emotion *topics* containing single word tokens are identified more easily than multi word *topics*. Sometimes, the emotion related *topics* coexist with other potential non-emotional *topics*. As the *topics* may consist of multi-word strings, the text spans denoting the *topic* spans create problems in identifying emotion *topic* span from other non-emotional *topic* spans. In the following example, the *emotional expression* আনন্দ *ananda* 'enjoy' is related to the topic গান *gan* 'song' and টিভি *TV* 'television'. The baseline system additionally captures বই *boi* 'book' that is a potential but non-emotion *topic*.

তুমি তো বই পড়তেই না, এখন দেখছি তুমি **গান**, **টিভি** তেও
(tumi) (to) (boi) (portei) (na), (ekhon) (dekhchi) (tumi) (**gan**), (**TV**) (teo)
আনন্দ পাওনা।
(*ananda*) (paona)
*You never used to read books; now we notice that you also don't enjoy **song/ television**.*

Solution: The *topic* of an opinion depends on the context in which its associated *opinion expression* occurs (Stoyanov & Cardie, 2008a). The *common rhetoric similarity* feature helps the syntactic system by aiming to separate emotion *topics* from non-emotion *topics* and to separate the overlapping possibilities of discrete emotion topic spans from non-topical contiguous regions. If the identified *topic* chunks are tagged with *common rhetoric similarity*, the chunks are classified as emotional *topics* and separated from non-topical elements in a sentence. The improvements at some important steps by incorporating the rule based post-processing techniques are shown in Table 2. It is observed that the simple rules have substantially reduced errors and have improved the performance of the system satisfactorily. The application of the post-processing techniques also achieves an alpha score of 0.6721 on the test set.

Table 2. The alpha scores of the system after handling the four error cases.

Cases	Krippendorff's α
Before Error Analysis	0.6332
Case 1	0.6476
Case 2	0.6417
Case 3	0.6498
Case 4	0.6402
Case 1+ Case 2	0.6510
Case 1+Case 3	0.6533
Case 1+Case 2+Case 3	0.6601
Case 1+Case 2+Case 4	0.6625
Case 1+Case 3+Case 4	0.6678
Case 1+Case 2+Case 3+Case 4	**0.6772**

7. Conclusion

The automatic extraction of *emotional expressions*, sentential emotion *holders*, and *topics* from Bengali blog data is accomplished in the present task. The supervised implementation of

the system shows improvement over the rule-based baseline because the rule-based system fails to capture the implicit textual clues whereas the supervised system captures the clues in terms of combined features. The evaluation of the co-reference using Krippendorff's alpha is helpful in diagnosing the importance of the three emotional components. The rule-based post-processing techniques for reducing the error cases have shown substantial improvement in the performance of the system. From the overall analysis, it is observed that the identification of emotional co-reference is helpful in identifying user-topic relations. The handling of metaphors and their impact in detecting sentence level emotion is not considered. Future analysis concerning the time based emotional change can be used for *topic* model representation. The need for co-reference requires that the presence of indirect affective clues can also be traced with the help of the *holder* and *topic*.

Reference

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. *HLT- EMNLP*, 579-586.

Banerjee, S., Das, D., & Bandyopadhyay, S. (2009). Bengali Verb Subcategorization Frame Acquisition - A Baseline Model. *ACL-IJCNLP-2009*, *ALR-7 Workshop,* 76-83.

Banerjee, S., Das, D., & Bandyopadhyay, S. (2010). Classification of Verbs – Towards Developing a Bengali Verb Subcategorization Lexicon. *GWC*, 76-83.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., & Jurafsky, D. (2004). Automatic Extraction of Opinion Propositions and their Holders, In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *Proceedings of HLT/EMNLP*.

Das, D., & Bandyopadhyay, S. (2009a). Word to Sentence Level Emotion Tagging for Bengali Blogs. *ACL-IJCNLP 2009,* 149-152.

Das, D., & Bandyopadhyay, S. (2009b). Emotion Tagging – A Comparative Study on Bengali and English Blogs. *ICON-09*. 177-184.

Das, D., & Bandyopadhyay, S. (2010a). Labeling Emotion in Bengali Blog Corpus – A Fine Grained Tagging at Sentence Level. *ALR8, COLING-2010*, 47-55.

Das, D., & Bandyopadhyay, S. (2010b). Developing Bengali WordNet Affect for Analyzing Emotion. *ICCPOL-2010*, 35-40.

Das, D., & Bandyopadhyay, S. (2010c). Sentence Level Emotion Tagging on Blog and News Corpora. *Journal of Intelligent System (JIS),* 19 (2), 125-134.

Das, D., & Bandyopadhyay, S. (2010d). Emotion Holder for Emotional Verbs – The role of Subject and Syntax. In *CICLing*, A. Gelbukh (Ed.), LNCS 6008, 385-393.

Ekbal, A., & Bandyopadhyay, S. (2008). Named Entity Recognition using Appropriate Unlabeled Data, Post-processing and Voting. In *Informatica Journal of Computing and Informatics*, ACTA Press.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, 48(4), 384-392.

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, *Language Resource and Evaluation Campaign*.

Evans, D. K. (2007). A low-resources approach to Opinion Analysis: Machine Learning and Simple Approaches, *NTCIR*.

Hu, J., Guan, C., Wang, M., & Lin, F. (2006). Model of Emotional Holder. *In Shi, Z.-Z., Sadananda, R. (eds.) PRIMA 2006. LNCS (LNAI)*, 4088, 534-539.

Joachims, T. (1998). Text Categorization with Support Machines: Learning with Many Relevant Features. In *European Conference on Machine Learning*, 137-142.

Kim, Y., Jung, Y., & Myaeng, S.-H. (2007). Identifying Opinion Holders in Opinion Text from Online Newspapers. In *2007 IEEE International Conference on Granular Computing*, 699-702, doi:10.1109/GrC.2007.45.

Kim, S. M., & Hovy, E. (2006). Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K., & Fukushima, T. (2004). Collecting evaluative expressions for opinion extraction. *IJCNLP*.

Kipper-Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania.

Krippendorff, K. (2004). Content analysis: An introduction to its methodology. *Thousand Oaks, CA: Sage*.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, *TEXT* 8, 243-281.

Mishne, G. & Rijke, de M. (2006). Capturing Global Mood Levels using Blog Posts. In *Proceedings of AAAI, Spring Symposium on Computational Approaches to Analysing Weblogs*, 145-152.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2007). Narrowing the Social Gap among People Involved in Global Dialog: Automatic Emotion Detection in Blog Posts, *ICWSM*.

Ng, V., & Cardie, C. (2002). Improving machine learning approaches to co-reference resolution. In *Proceedings of ACL*.

Popescu, A., & Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP*.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1-47.

Seki, Y. (2007). Opinion Holder Extraction from Author and Authority Viewpoints. In *Proceedings of the SIGIR'07*, *ACM 978-1-59593-597-7/07/0007*.

Soon, W., Ng, H., & Lim, D. (2001). A machine learning approach to co-reference resolution of noun phrases. *Computational Linguistics*, 27(4), 521-544.

Stoyanov, V., & Cardie, C. (2008a). Annotating topics of opinions. In *Proceedings of Language Resource and Evaluation Campaign*.

Stoyanov, V., & Cardie, C. (2008b). Topic Identification for Fine-Grained Opinion Analysis. *Coling 2008*, 817-824.

Strapparava, C., & Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text, *ACL*.

Swier, R. S., & Stevenson, S. (2004). Unsupervised Semantic Role Labelling. *EMNLP*.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2), 1-54.

Yang, C., Lin, K. H.-Y., & Chen, H.-H. (2007). Emotion classification Using Web Blog Corpora. In *Proceedings of the IEEE, WIC, ACM International Conference on Web Intelligence*, 275-278.

Yang, C., Lin, K. H.-Y., & Chen, H. H. (2009). Writer Meets Reader: Emotion Analysis of Social Media from both the Writer's and Reader's Perspectives. In *Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 287-290.

Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of ICDM*.

Zhang, Y., Li, Z., Ren, F., & Kuroiwa, S. (2008). A preliminary research of Chinese emotion classification model. *IJCSNS*, 8(11), 127-132.

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:_____(M/Y)

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

CARD HOLDER SIGNATURE: _____

Phone No.: _____E-mail: _____

Address: _____

PAYMENT FOR

US$ ____ ❑ Computational Linguistics & Chinese Languages Processing (IJCLCLP)

Quantity Wanted: _____

US$ ____ ❑ Journal of Information Science and Engineering (JISE)

Quantity Wanted: _____

US$ ____ ❑ Publications:_____

US$ ____ ❑ Text Corpora: _____

US$ ____ ❑ Speech Corpora:_____

US$ ____ ❑ Others: _____

US$ _____ ❑ Membership Fees ❑ Life Membership ❑ New Membership ❑Renew

US$ _____ = Total

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

ACLCLP
% IIS, Academia Sinica
Rm502, No.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：_____

信用卡號：_____-_____-_____-_____　有效日期：_____(m/y)

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$_____ ❑ Journal of Information Science and Engineering (JISE)

NT$_____ ❑ 中研院詞庫小組技術報告_____

NT$_____ ❑ 文字語料庫 _____

NT$_____ ❑ 語音資料庫 _____

NT$_____ ❑ 光華雜誌語料庫1976~2010

NT$_____ ❑ 中文資訊檢索標竿測試集/文件集

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:

11529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.18 no.1 March 2013

Contents

Papers