Sina Mandarin Alphabetical Words: A Web-driven Code-mixing Lexical Resource

Rong Xiang[1], Mingyu Wan[12], Qi Su[2], Chu-Ren Huang[1], Qin Lu[1]
[1]The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hong Kong (China)
csrxiang,csluqin@comp.polyu.edu.hk, churen.huang@polyu.edu.hk
[2]Peking University, 5 Yiheyuan Road, Beijing (China)
wanmy, sukia@pku.edu.cn

Abstract

Mandarin Alphabetical Word (MAW) is one indispensable component of Modern Chinese that demonstrates unique code-mixing idiosyncrasies influenced by language exchanges. Yet, this interesting phenomenon has not been properly addressed and is mostly excluded from the Chinese language system. This paper addresses the core problem of MAW identification and proposes to construct a large collection of MAWs from Sina Weibo (SMAW) using an automatic web-based technique which includes rule-based identification, informatics-based extraction, as well as Baidu search engine validation. A collection of 16,207 qualified SMAWs are obtained using this technique along with an annotated corpus of more than 200,000 sentences for linguistic research and applicable inquiries.

1 Introduction

Mandarin Alphabetic Words (MAWs), also known as lettered words (Liu, 1994) or code-mixing words (Nguyen and Cornips, 2016), are usually formed by the 26 Latin alphabets in combination with Chinese characters, e.g. "X-光/X射线" (X-ray). Although pure alphabets (e.g. "NBA") used in the Chinese context have also been regarded as MAWs (Liu, 1994; Huang and Liu, 2017), they are more like switching-codes that retain the orthography and linguistic behaviors of the original language, instead of showing typical Chinese lexical characteristics. It is noteworthy that MAWs shall be taken as a code-mixing phenomenon instead of code-switching as a MAW is still a Chinese word which is not switched into another language. Therefore, in this work, MAWS refer to the combined type which encodes both alphabet(s) and Chinese character(s) in one word, such as "A型"(A-type), "PO主" (post owner), and "维生素C/维C" (Vitamin C).

It is linguistically-interesting and applicably-significant to investigate ~~the combined type of~~ MAWs due to ~~the following~~ two main reasons. First, A MAW maintains part of the Chinese characteristics in morphology, phonology and orthography (e.g. "PK-过" (player killed, past tense)), and in the meanwhile it also demonstrates some properties of the foreigner language (e.g. "维生素-*ing*" (supplementing Vitamin, progressive)), providing a unique lexical resource for studying the morpho-phonological idiosyncrasies of code-mixing words. Second, it serves as an indispensable part of people's daily vocabulary, especially under the rapid development of social media communication, yet being out-liars of the Chinese lexicon, causing problems to existing word segmentation/new word extraction tools that are trained on traditional words (Chen and Liu, 1992; Xue and Shen, 2003).

Consider the following example:

E1:	**PO**主也不知道链接被吞了
	(The post owner didn't know that the link has been hacked off)
Seg:	**PO**/主/也/不/知道/链接/被/吞/了
Golden Seg:	**PO**主/也/不/知道/链接/被/吞/了

The sentence in E1 (example 1) is segmented using Stanford Parser (Manning et al., 2014) which fails to identify the word "PO主" and breaks it into two parts. The same type of error also occurs in other renowned segmentation tools. Although Huang et al. (2007) proposed a radical method of word segmentation to meet the challenge, using a concept of classifying a string of character-boundaries (CB's) into either word-boundaries (WB's) or non-word-boundaries, their work did not address the cases of code-mixing words, whose word boundaries can also fall on the alphabets. Some other methods mainly rely on unsupervised methods (Chang and Su, 1997) or simple statistical

methods by focusing on N-gram frequencies, with indexes of collocation and co-occurrence (Chang and Su, 1997; Chen and Ma, 2002; Dias, 2003). However, these works are mainly designed for new words of pure Chinese characters, which are not applicable to alphabetical words.

In this paper, we address the issue of MAW identification and present the construction of the **Sina MAW (SMAW) lexicon**. (available at https://github.com/Christainx/SMAW) using a fully automatic information extraction technique. The quality of the MAWs (accurateness and inter-rater agreement) are rated by three knowledge experts for system evaluation. Compared to previous related resources, the current data provides an unprecedentedly large, balanced, and structured MAWs that are popularly used in Sina Weibo after 2010s. It is reasonable to anticipate that, with more MAWs being included in the Chinese lexicon, it shall benefit many Chinese language processing tasks which need to deal with code-mixing word segmentation.

2 Related Works

The earliest MAW was probably "X射线/X-光" (X-ray), which was officially documented in 1903 (Zhang, 2005). For over 60 years, such words had been largely confined to technical and medical domains with very few lexicalized and registered terms in dictionaries. The authoritative *Xiandai Hanyu Cidian/XianHan ("现代汉语词典")*, for instance, initiated a separate section to include 39 MAW entries in 1996. This list has grown rapidly with each subsequent XianHan dictionary edition, reaching 239 entries by the 2012 edition. This in turn generated a flurry of related linguistic studies, which were mainly focused on lexicological and language policy issues (Su and Wu, 2013; Zhang, 2013). Some works have dealt with the emergence of MAWs in light of globalization, placing them in a socio-cultural context (Kozha, 2012; Miao, 2005), and a few are also interested in studying the morpho-lexical status of MAWs (Lun, 2013; Riha and Baker, 2010; Riha, 2010).

In the age of Internet and social media, the scale of MAW, their extraction methods, and resources of MAWs have changed drastically since the last decade. For example, Zheng et al. (2005) extracted a small set of MAWs with manual validation from the corpus of People's Daily (Year 2002). Jiang and Dang (2007) extracted 93 MAWs (out of 1,053 new domain-specific terms) using a statistical approach with rule-based validation. Recently, Huang and Liu (2017) extracted over 1,157 MAWs from both the Sinica Corpus (Chen et al., 1996) and the Chinese Gigaword Corpus (Huang, 2009) based on manually segmented MAWs in the corpora. Although they have extracted 60,000 MAWs, but the list mainly includes pure alphabets those are indeed switching codes of other languages, instead of real MAWs. Their work has established a taxonomy of distributional patterns of alphabetical letters in MAWs and found that typical MAWs follow Chinese modifier-modified (head) morphological rule and the most frequent and productive pattern is alphabetical letter+ mandarin character (AC), such as the type B in the form of "B型".

Despite of the above investigations, MAWs have not been identified in a systemic and automatic way. The problem of identifying MAWs can be generalized as an issue of new/unknown/out-of-vocabulary word extraction (code-mixing Chinese words in particular) (Chen and Ma, 2002; Zhang et al., 2010). A commonly adopted way of identifying a new word usually rely on word segmentation at the first step and then map the candidates to an existing dictionary. Those not mapped in the dictionary will be identified as new words. This is actually problematic for identifying MAWs (cf. example in Section 1). In addition, the previous studies mainly adopt manual selection for MAW construction from newspapers of pre-1990s (Huang and Liu, 2017), hence the resources are domain-constrained and usage-outdated. To complement such a bias, we propose to collect social-media style MAWs that are commonly used in Sina Weibo, a near-natural context. Moreover, as there are many debates among linguists about the definition of a MAW (Ding et al., 2017; Liu, 1994; Tan et al., 2005; Xue, 2007; Liu, 2002), the current work utilizes data-driven statistical approach as well as leveraging on search engine hits to exclude pseudo-MAWs of low-vitality. Details of the methodology are given in the next section.

3 Construction of SMAW

The construction of the SMAW dataset is carried out through an alphabet-anchored brute-force extraction of n-grams (n = 5 window size) at the initial stage, which creates substantial false negative candidates, such as sub-component of a positive case ("啦A梦", Doraemon), a long-distance

segmentation (“A股反弹”, rally of Shanghai SE Composite Index), etc. To eliminate as many negative cases as possible, we implement the following three phases of candidate filtering: rule-based refinement, informatics-based extraction, and search engine validation, as shown in Figure 1.

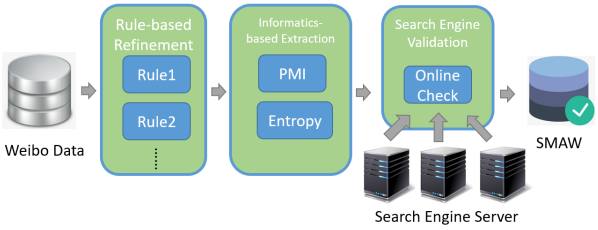

Figure 1: The framework of the SMAW construction

A number of selected rules are set as a preliminary refinement. These rules are easy implemented and fast executed. PMI (Point-wise Mutual Information) and entropy are calculated for selecting components of high co-occurrence rate and informative flexibility. Using informatics-based methods can greatly help narrow down the scope of MAW candidates. Last, search engine is adopted to filter out low-vitality terms based on user links. This intellectual agent provide use cases about a candidate word as extra evidence. Details of these steps are described in the following subsections.

3.1 Rule-based Refinement

The data source for the extraction of MAW candidates is Sina Weibo (Weibo for short, or micro-blogs). Weibo is one of the most popular social media platforms in China with over 400 million active users per month. All registered users with internet access can publish, re-post, comment or issue "like" for any post at no cost. This platform becomes the enabler for generating tons of data online, which can serve as a huge Web corpus. The raw dataset crawled from Weibo consists of over 226 million posts (around 20 gigabytes data). Below is an typical example of a user post in this dataset which includes a number of web-specific linguistic usages.

E2: #BMW赛车纪录片#
#亚洲公路摩托锦标赛珠海站全记录#
@UNIQ-王一博http://t.cn/EPdahkI
(#BMW Racing Documentary#Records Zhuhai
(in Asian Highway Motorcycle Championship.
@AX12FZ32 http://t.cn/EPdahkI)

As shown in E2, among all alphabetical chunks, many candidates are URL links or related to topics (surrounded by #) or user names (introduced by the “@” symbol). These alphabetical sequences are confirmed as noises that could be readily excluded from the final data. Besides the above noises, there are some other false MAW candidates that demonstrate obvious patterns. For example, candidates of emoji (e.g. “QAQ”, “LOL”, “:P”, “T_T”) are transformed symbols that encode no lexical meanings and shall be eliminated from the MAW list. Using rule-based filtering of the above unambiguous noises can largely reduce the sample size and retain the high coverage of the lexicon. The information about the elimination steps and data sizes is provided in Table 1.

3.2 Informatics-based Extraction

After Rule-based Refinement, a set of 1,470k potential MAW candidates is obtained. Term-frequency (TF) is a commonly adopted metric to filter out low-occurrence candidates. However, it is insufficient to identify MAW only by TF. For instance, both “A股” (Shanghai SE Composite Index) and “A股反弹” (rally of Shanghai SE Composite Index) have high TF but only “A股” is a valid MAW. In this work, informatics-based methods are used to automatically filter the negative cases, including PMI for measuring the internal cohesion, and entropy for measuring the external uncertainty of the candidates.

Pointwise mutual information (PMI) is proposed by Bouma (2009) to measure the probability of a particular co-occurrence of two variables. Let w be an MAW candidate that consists of two components c_1, c_2. The PMI of w can be calculated via Formula 1.

$$PMI(c_1;c_2) = -log(\frac{p(c_1,c_2)}{p(c_1)*p(c_2)}) \quad (1)$$

In practice, at least one component, denoted as c_a must contain alphabet character(s). If w consists of more than three components, we use the combination coordinated by c_a. For example, “哆啦, A, 梦” (Doraemon) can be computed by using “哆啦A, 梦” and “哆啦, A梦”. Then, Formula 1 can be adapted to Formula 2 in this circumstance.

$$PMI(w) = min(PMI(c_1;c_a), PMI(c_a;c_2)) \quad (2)$$

The threshold of PMI is experimentally set to -16.2. In addition to the above metric of measuring

Rule	Description	Quantity
NONE	brute force candidates collection	25,594k
Topic	remove candidates with '#'	165k
Username	remove candidates with '@'	297k
No Chinese	remove candidates without Chinese character	1,302k
Too Short Length	remove candidates less than LEN_THRES characters	595k
Rare Occurance	remove candidates which count less than FREQ_THRES	18,443k
English Expression	remove candidates contain two or more English words	1,421k
Symbol	remove candidates contain symbols such as '&' and '*'	419k
Emoji	remove candidates contain emoji such as "XDD"	193k
POS tag	remove candidates with invalid POS tag such as 'DET'	1k
ALL RULES	using all rule-based refinement	1,470k

Table 1: Rule-based Refinement and Data Sizes.

the internal "fixedness" of a word, another dimension for identifying word boundaries is to refer to information entropy of its collocation environment. As proposed by He and Jun-Fang (2006), information entropy can be used to measure the uncertainty (flexibility) of a candidate's environment, the larger the more flexible, and the more likely the candidate being a word. Consider the negative case of "素C" which only occurs in the context of "维生素C" (Vitamin C) (entropy in this case is low). In contrast, the positive case "维生素C" occur in many different contexts: 补充"维生素C" (Take Vitamin C), 高剂量"维生素C" (High-dosage Vitamin C), "维生素C"对感冒有效(Vitamin C copes with colds), etc. (entropy in this case is high). Let c_h and c_t be the head and tail components surrounding w. The head entropy of w, as an example, is generated by Formula 3. And the final entropy of w is obtained by $min(H(h), H(t))$.

$$H(h) = -\sum p(c_h)_i * log(p(c_h)_i) \quad (3)$$

The upper bound of entropy is determined to 0.2 in our implementation. By using PMI and entropy, we have eliminated 878k and 560k invalid MAW candidates respectively, having 33k candidates remain in the list.

3.3 Search Engine Validation

This section aims to further filter out pseudo MAWs which are either less frequently used or look like words in the form but barely carry lexical meanings. Search engine such as Google, Bing and Baidu provides a superb knowledge base to reflect the semantic information of a MAW candidate, as active MAW candidates with more links are more likely to carry authentic meanings. On the other hand, semantic information can help to exclude non-lexicon candidates. For instance, "UNIQ-王一博" is a famous Chinese actor in the band "UNIQ". The features of this false candidate can perfectly pass previous filtering methods. Thus, more intellectual validation schema is demanded. In current implementation, Baidu, the most popular search engine in China, is adopted as the knowledge agent for retrieving the vitality evidence of the candidates. Figure 2 is the flowchart of search engine filtering.

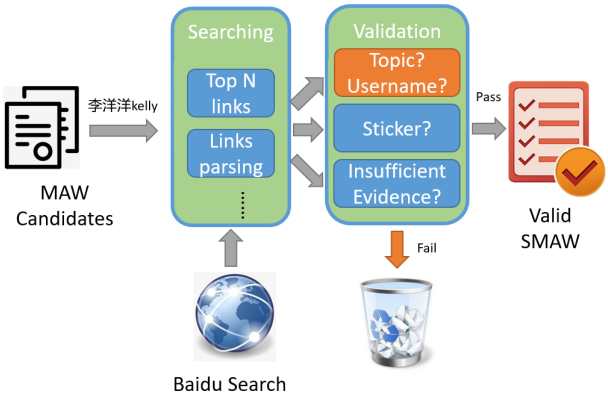

Figure 2: Flowchart of Search Engine Validation

We use the example of "李洋洋kelly" (a username combined with a Chinese name and an English nickname) for illustration. The top N links (10 by default) are first collected as external evidence. The linked text is then cleaned and parsed to check whether one MAW candidate exists. For instance, the term "李洋洋kelly" occurs only in cases of "@李洋洋kelly", so it is identified as a username, which is an invalid MAW. Except for username checking, stickers and insufficient evidence are also utilized for validation. In our proposed model, the threshold of links is configured

as 5.

4 Results and Evaluation

After the above procedure, we have finally collected a set of 16,207 SMAWs, as well as an annotated corpus of more than 200,000 sentences containing this SMAWs. Evaluation on the quality of the lexicon and details of this resource are given in the following sections.

4.1 Evaluation

This section aims to estimate the performance of the proposed extraction method in terms of **Accuracy** and **Inter-rater agreement** by human raters. As MAWs demonstrate a dynamic role in the Chinese lexicon, it is unable to refer to a full reference set for calculating Recall and Precision. Instead, we use accuracy to measure its quality.

In the evaluation, three groups of SMAWs (100 each group, 300 in total) are randomly sampled from each phase for the participants to judge the acceptance of the candidates. Raters are asked to make judgements of 1 if they think the candidate is a MAW, or 0 vice versa. Then, Accuracy (Acc.) is calculated as the average of the three groups' acceptance rates.

Inter-rater agreement among the three raters is also measured using the Cohen's Kappa Coefficient (*K.*) (Kraemer, 2014). Cohens Kappa was first introduced as a measure of agreement between observers of psychological behavior to measure the degree of agreement or disagreement of two or more people observing the same phenomenon. It is applicable to the measurement of the performance of the current approach. The evaluation results are given in Table 2.

Phase	Metric	Acc.	δ^1	*K.*
1	TF	.12	-.03	.56
2	+Rule-based	.22	.10	.58
3	+PMI	.62	**.40**	.65
4	+Entropy	.77	.15	.70
5	+Baidu	.82	.05	**.78**
B0	TF+Max.	.15	-	.59

Table 2: The Evaluation Results

Table 2 summarizes the accumulative performances of using the various metrics for candidate selection during each phase. B0 is a baseline method that simply employs term frequency and the maximal sequence principle. For example, using the maximal sequence principle will select "哆啦A梦" (Doraemon) over candidates of "啦A梦" or "A梦". However, this method could also make false decisions, such as in "安全使用免费WiFi" (Safely use free wifi) where "免费WiFi" (free wifi) shall be a positive candidate.

The current work makes use of a more reliable extraction approach that is proved to be obviously more effective for the identification of alphabetical words (Acc. = 0.82, *K.* = 0.78). The PMI metric contributes the most to the improvement of Accuracy (0.40), followed by Entropy (0.15), indicating the usefulness of informatics-based metrics for word identification. In addition, the incremental *K.* of each phase suggests the increased agreement among the three raters by adopting the several metrics, especially after the Baidu search engine validation.

Top 50 MAWs in Web

Top 50 MAWs in Giga

Figure 3: Word clouds of MAWs in Web and Giga

[1]The accuracy discrepancy between the current metric and the baseline

The high accuracy score and agreement in the evaluation has proven the effectiveness of the extraction method, as well as demonstrating a good quality of the lexicon.

4.2 The Lexical Characteristics

This section will analyse the lexical properties of the SMAW lexicon. Comparisons between the SMAW list ("Web" hereinafter) and the MAWs in Huang and Liu (2017) ("Giga" hereinafter) will be made in terms of key vocabulary, length distribution, word formation types and lexical diversity so as to highlight the lexical differences of MAW between newspaper and social media, as well as reflecting the lexical development of alphabetical words in the recent two decades.

4.2.1 Vocabulary

To have a glance at the vocabularies of the two lexicons, top 50 MAWs of each set are demonstrated in Figure 3. The size of the words indicate its usage frequency.

It can be observed that the most frequent MAW in the Giga list is "B型" (B-type), while in the Web list, the most frequent MAW is "HOLD住" (To endure), which is a typical Internet neology. Moreover, most MAWs in Giga are disyllabic, e.g. "A型" (A-type) and "A级"(A-level), while SMAWs tend to be more lengthy, containing words of a wider range of syllables (e.g. "NBA全明星" (NBA all-star)). Specifically, MAWs in Giga show a dominant (rigid) pattern of "X类/型" (Type-X). However, in Web, MAWs refer to more Part-of-Speech categories, including verbs (e.g. "Hold住"), nouns (e.g. "BB霜" (BB cream)), or adjectives (e.g. "牛X" (incredibly awesome)), indicating the trend of MAWs accounting for more grammatical roles in the Chinese language. Lastly, the lexical senses of Giga MAWs are more concentrated to the "type/classification" meaning, while MAWs in Web encode a wider range of meanings, including name entities, swear words, economics, entertainment, etc.

The above keyword differences reflect a dramatic change of MAWs at syllabic, lexical, grammatical and semantic levels in the recent decades.

4.2.2 Length Distribution

The box-plots in Figure 4 provide an overview of the length distribution of MAWs in Giga (Huang and Liu, 2017) and Web (SMAW).

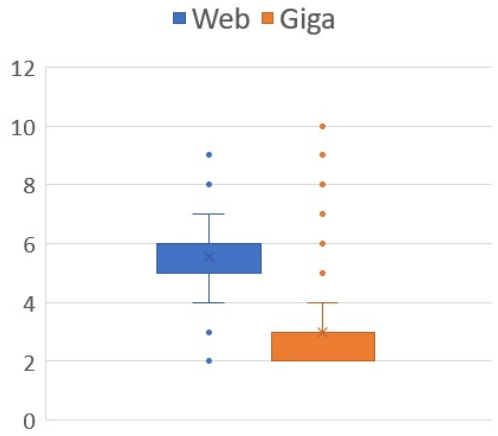

Figure 4: Length distribution of MAWs in Giga and Web

As shown in Figure 4, MAWs in Web are much longer and more scattered than that in Giga. The mean length of MAWs in Giga is 2-3, but in Web the mean length is around 5. Overall, the MAWs in Web are distributed across a wider span. This may imply a tendency of code-mixing words being longer and richer in Modern Chinese lexicon.

4.2.3 Word Formation

In line with the work of Huang and Liu (2017), word formation of the MAWs is classified into four major types according to the positions of the A (alphabet) and C (character), including AC (e.g. "x-光"), CA (e.g. "牛b"), CAC (e.g. "程I青" (a Chinese Name)) and other types. The number of the four types of MAWs in Giga and Web is displayed in Table 3 for comparison.

	AC	CA	CAC	Other	Total
Giga	665	283	185	18	1151
(pct)	**57.8%**	24.6%	16.1%	1.5%	100.0%
Web	6971	6994	2242	0	16207
(pct)	43.0%	**43.2%**	13.8%	0.0%	100.0%

Table 3: Word formation comparison

As highlighted in Table 3, the dominant type in Giga is AC, while CA is more prevalent in Web. Huang and Liu (2017) explained to the dominance of AC type with the modifier-modified compound structure in Chinese because heads of nouns are usually right positioned (Sun, 2006). However, MAWs in Web account for a wider grammatical roles and more verbs are found in the new list. Contrary to nouns, verbs are left headed, such as in "打call" (cheer up), where "打" (beat) is the head. In addition, cases like "维c" (Vitamin C),

"双c" (double cores), and "最In" (Most popular) are headed on alphabets instead of the Chinese character, indicating that heads are not necessarily positioned at the Chinese characters.

4.2.4 Lexical Diversity

TTR (type–token ratio) is used to measure the lexical diversity/richness of a language (Durán et al., 2004). This metric is adopted here, with data normalized (STTR), for measuring the lexical diversity of the MAWs in Giga and Web, as shown in Table 4.

Data	STTR	AC.STTR	CA.STTR
Web	14.53	16.9	12.3
Giga	8.77	7.6	15.2

Table 4: Lexical Diversity Comparison

Table 4 seems to suggest a reverse relation between the frequency of the MAW types and their lexical richness: the "AC" type is dominant in Giga, but it demonstrates a lower STTR; similarly, the "CA" type is dominant in Web, and it also shows a lower STTR. Overall, the Web MAWs show a richer vocabulary compared to the newspaper MAWs (Giga), indicating the higher productivity of social media language.

4.3 The Corpus

In addition to the SMAW lexicon, we have also retrieved more than 200,000 sentences (around 2,000,000 tokens) for the 16,207 SMAW (each SMAW contains 10 or so sentences) to construct a SMAW corpus which can support code-mixing words inquiries.

The characters in the sentences are all transferred into simplified Chinese for consistency. The sentences were automatically segmented using Stanford CoreNLP [2](Manning et al., 2014). The automatic word segmentation is enabled as the alphabetical words are pre-identified in our SMAW lexicon. With confirmed boundaries of the alphabetical words, it becomes an ordinary task of segmenting the remaining Chinese characters.

On the basis of the raw sentences, we are building a concordance engine for loading the content of the corpus following the Chinese Word Sketch schema (Hong and Huang, 2006), which can support users' inquires of word and grammatical collocations of code-mixing words. Samples of the corpus are shown in the following interface.

一定(D) 要(D)	**HOLD住(VA)**	！
疯狂(D) 店庆(VA) 11天(Nd)，还(D) 能(D)	**HOLD住(VA)**	吗(T)
KITTY控(Na) 们(Na) 还(D)	**HOLD住(VA)**	吗(T)
微时代(Na) ， 大(A) 趋势(Na) ， 可得(VH)	**HOLD住(VA)**	！
亲(I) ！ 你(Nh) 要(D)	**HOLD住(VA)**	哦(T)
大家(Nh)	**HOLD住(VA)**	哦(T)
各位(Nes) 看官(Na) 要(D)	**HOLD住(VA)**	了(Di)

Interface 1: Corpus samples of "HOLD住" (KWIC)

Besides, the corpus is undergoing a POS tagging process using the Academia Sinica segmentation and tagging system (Chen et al., 1996; Zhao et al., 2006) in order to support grammatical inquiries of linguistic accounts. The tagging has been performed automatically with manual post-checking on the SMAWs. The precision accuracy is estimated to be over 85%. Since the tagging is still in progress, we provide the POS distribution [3] of most frequent 50 SMAWs to give a general view of the grammatical distribution of popular SMAWs. Figure 5 below shows the POS distribution of MAWs in Web and Giga for comparison purpose.

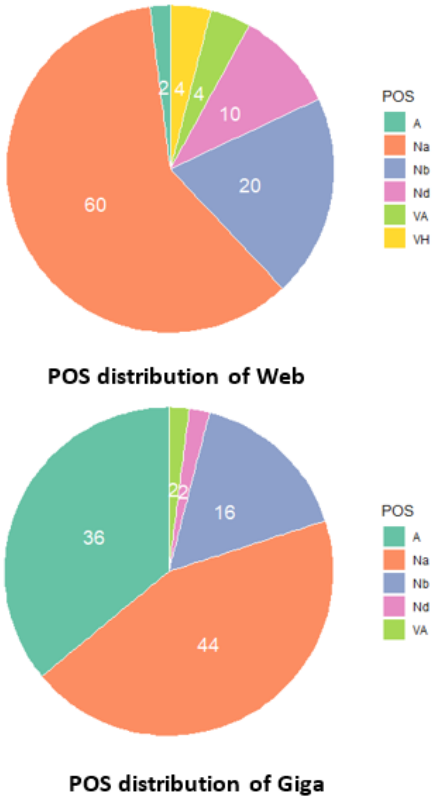

Figure 5: POS distribution of MAWs in Giga and Web

[2]https://stanfordnlp.github.io/CoreNLP/

[3]https://catalog.ldc.upenn.edu/LDC2009T14

The POS distribution in Figure 5 implies that MAWs has developed a more salient role in the Chinese lexicon: from mainly nouns (Na, Nb, Nd) to verbs (VA, VH), from modifiers (A) to core lexical components (heads and arguments), and the graph demonstrates a more diversified lexcial categories (more divisions and colorful) of new MAWs.

5 Conclusion and Future Work

This work utilizes social media (Sina Weibo) and search engine (Baidu) for collection and validation of code-mixing words to tackle the under-representation and identification problems of MAWs. The evaluation of the current MAW dataset proves the high performance (Acc. = 0.82, *K.* = 0.78) of the proposed extraction method, showing the usefulness of informatics-based metrics for word identification. The contribution of this work is two-fold: it proposes an innovative method of leveraging the Web for MAW extraction without involvement of manual mediation, yet achieving promising performance in identifying out-of-vocabulary code-mixing words; it provides a unique MAW dataset that is most updated, scaled, structured and comprehensive for supporting linguistic inquiries of code-mixing words, as well as for facilitating related NLP tasks. The preliminary analysis to the lexical and grammatical characteristics of SMAWs and the corpus implies the development of code-mixing words into being a more important and diversified component in the Chinese lexicon. Future work will continue the annotation of the lexicon and the corpus with information of domains, sources, active time, semantic classes, etc., and conduct deeper linguistic analyses for uncovering the phonological and morpho-lexical characteristics of code-mixing words.

Acknowledgments

We acknowledge the research grants from Hong Kong Polytechnic University (PolyU RTVU) and GRF grant (CERG PolyU 15211/14E, PolyU 152006/16E).

This work is funded by the GRF grant (PolyU 156086/18H) and the Post-doctoral project (no. 4-ZZKE) at the Hong Kong Polytechnic University.

References

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Jing-Shin Chang and Keh-Yih Su. 1997. An unsupervised iterative method for chinese new lexicon extraction. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 2, Number 2, August 1997*, pages 97–148.

Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. 1996. Sinica corpus: Design methodology for balanced corpora. In *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation*, pages 167–176.

Keh-Jiann Chen and Shing-Huan Liu. 1992. Word identification for mandarin chinese sentences. In *Proceedings of the 14th conference on Computational linguistics-Volume 1*, pages 101–107. Association for Computational Linguistics.

Keh-Jiann Chen and Wei-Yun Ma. 2002. Unknown word extraction for chinese documents. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Gaël Dias. 2003. Multiword unit hybrid extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 41–48. Association for Computational Linguistics.

Hongwei Ding, Yuanyuan Zhang, Hongchao Liu, and Chu-Ren Huang. 2017. A preliminary phonetic investigation of alphabetic words in mandarin chinese. In *INTERSPEECH*, pages 3028–3032.

Pilar Durán, David Malvern, Brian Richards, and Ngoni Chipere. 2004. Developmental trends in lexical diversity. *Applied Linguistics*, 25(2):220–242.

Jia-Fei Hong and Chu-Ren Huang. 2006. Using chinese gigaword corpus and chinese word sketch in linguistic research. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pages 183–190.

Chu-Ren Huang. 2009. Tagged chinese gigaword version 2.0, ldc2009t14. *Linguistic Data Consortium*.

Chu-Ren Huang and Hongchao Liu. 2017. Corpus-based automatic extraction and analysis of mandarin alphabetic words (in chinese). *Journal of Yunnan Teachers University. Philosophy and social science section*.

Chu-Ren Huang, Petr Šimon, Shu-Kai Hsieh, and Laurent Prévot. 2007. Rethinking chinese word segmentation: tokenization, character classification, or wordbreak identification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 69–72.

Shaohua Jiang and Yanzhong Dang. 2007. Automatic extraction of new-domain terms containing chinese lettered words (in chinese). *Computing Engineering*, 33(2):47–49.

Ksenia Kozha. 2012. Chinese via english: A case study of "lettered-words" as a way of integration into global communication. In *Chinese Under Globalization: Emerging Trends in Language Use in China*, pages 105–125. World Scientific.

Helena C Kraemer. 2014. Kappa coefficient. *Wiley StatsRef: Statistics Reference Online*, pages 1–4.

Yongquan Liu. 1994. Survey on chinese lettered words (in chinese). *Language Planning*, (10):7–9.

Yongquan Liu. 2002. The issue of lettered words in chinese. *Applied Linguistics*, 1:8S–90.

Ka Yee Lun. 2013. Morphological structure of the chinese lettered words. *University of Washington Working Papers in Linguistics*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Ruiqin Miao. 2005. *Loanword adaptation in Mandarin Chinese: Perceptual, phonological and sociolinguistic factors*. Ph.D. thesis, Stony Brook University.

Dong Nguyen and Leonie Cornips. 2016. Automatic detection of intra-word code-switching. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 82–86.

He Ren and Jun-fang Zeng. 2006. A chinese word extraction algorithm based on information entropy. *Journal of Chinese Information Processing*, 20(5):40–43.

Helena Riha. 2010. Lettered words in chinese: Roman letters as morpheme-syllables.

Helena Riha and Kirk Baker. 2010. Using roman letters to create words in chinese. In *Variation and Change in Morphology: Selected papers from the 13th International Morphology Meeting, Vienna, February 2008*, volume 310, page 193. John Benjamins Publishing.

Xinchun Su and Xiaofang Wu. 2013. Vitality and limitation of chinese lettered words (in chinese). *Journal of Beihua University(Social Sciences)*, 2.

Chaofen Sun. 2006. *Chinese: A linguistic introduction*. Cambridge University Press.

Li Hai Tan, Angela R Laird, Karl Li, and Peter T Fox. 2005. Neuroanatomical correlates of phonological processing of chinese characters and alphabetic words: A meta-analysis. *Human brain mapping*, 25(1):83–91.

Nianwen Xue and Libin Shen. 2003. Chinese word segmentation as lmr tagging. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 176–179. Association for Computational Linguistics.

Xiaocong Xue. 2007. A review on studies of lettered-words in contemporary chinese. *Chinese Language Learning*, 2.

Haijun Zhang, Heyan Huang, Chaoyong Zhu, and Shumin Shi. 2010. A pragmatic model for new chinese word extraction. In *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010)*, pages 1–8. IEEE.

Tiewen Zhang. 2005. Study of the word family 'x-ray' in chinese (in chinese). *Terminology Standardization & Information Technology*, 1.

Tiewen Zhang. 2013. The use of chinese lettered-words is a normal phenomenon of language contact. (in chinese). *Journal of Beihua University(Social Sciences)*, 2.

Hai Zhao, Changning Huang, and Mu Li. 2006. An improved chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 162–165.

Zezhi Zheng, Pu Zhang, and Jianguo Yang. 2005. Corpus-based extraction of chinese lettered words (in chinese). *Journal of Chinese Information Processing*, 19(2):79–86.