Transfer of Frames from English FrameNet to Construct Chinese FrameNet: A Bilingual Corpus-based Approach

Tsung-Han Yang, Hen-Hsen Huang, An-Zi Yen, and Hsin-Hsi Chen

Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
{thyang, hhhuang, azyen}@nlg.csie.ntu.edu.tw, hhchen@ntu.edu.tw

Abstract

Current publicly available Chinese FrameNet has a relatively low coverage of frames and lexical units compared with FrameNet in other languages. Frames are incompletely specified for some lexical units, and some critical lexical elements are even missing. That results in suitable frames cannot be triggered and filled in practical applications. This paper presents an automatic approach to constructing Chinese FrameNet. We first capture the mapping between English lexical entries and their Chinese counterparts in a large scale sentence-aligned English-Chinese bilingual corpus. Then, a semantic transfer approach is proposed based on word alignments applied to a large balanced bilingual corpus. The resource currently covers 779 frames and 36k lexical units. We apply it to annotate diary and tweet, and achieve overall 86% success rate to provide frame recommendations that are acceptable by annotators. The success rates in terms of source types are 95% and 80% for diaries and tweets respectively.

Keywords: FrameNet, Chinese, frame semantics, lexical units

1. Introduction

Semantic role labeling (SRL) is a fundamental task for many NLP applications. Given a context, SRL is aimed at identifying the semantic roles, or the set of semantic properties and relationships defined over a lexical unit (LU) or a target. The resources such as FrameNet (Baker et al., 1998; Fillmore et al., 2003) and PropBank (Kingsbury and Palmer, 2002) storing abundant information about lexical and predicate-argument semantics have advanced the field of semantic analysis further and made it possible for learning algorithms to train upon.

1.1 FrameNet

FrameNet is a lexical database with rich human-annotated semantic content based on the linguistic theory of Frame Semantics proposed by Fillmore (1982). FrameNet provides both human and machine readable structure for semantic frames, associated with frame elements (FEs), lexical units (LUs), and sample sentences.

The words that evoke a specific frame are called LUs or targets. Accompanied with each frame, there are a set of FEs defining semantic roles that is meaningful to that frame. Most importantly, FrameNet provides over 200,000 human annotated sentences associated to more than 1,200 semantic frames, and forms a great resource for analyzing semantic structures in natural language.

1.2 FrameNet in Chinese

Unlike English FrameNet, which has been in operation since 1997, Chinese FrameNet (You et al, 2005) started constructing the resource in 2005. To date, Chinese FrameNet (CFN) contains 3,947 LUs, 323 semantic frames, and 20,000 annotated sentences. Compared with English FrameNet, in which 13,638 LUs, 1,221 semantic frames, and 200,000 annotated sentences are provided, Chinese FrameNet is considerably smaller. In other words, Chinese FrameNet has a much low coverage of frames and lexical units, and results in limited applications.

1.3 FrameNet Construction in Other Languages

Efforts have been made to construct FrameNet resources in other languages. Most of which construct their resources with human annotation one by one laboriously, such as Japanese FrameNet (Ohara et al., 2003). Park et al. (2014) conduct the construction of Korean FrameNet by hiring trained translators to import 4,025 sentences selected from English FrameNet. Kim et al. (2016) further import additional 1,795 sentences to Korean FrameNet from Japanese FrameNet based on the similarities between these two languages. However, the high construction cost of such resources sometimes hinders it from growing to the proper scale that is applicable for real NLP tasks. Tonelli et al. (2008) propose an algorithm that projects English frames onto Italian ones, so that FrameNet in Italian could be constructed more easily.

In this work, we propose a novel approach to automate FrameNet construction. Based on a large-scale bilingual corpus, we transfer the machine-annotated FEs from English sentences to their Chinese counterparts based on the assumption of semantic equivalence between both source sentences and target sentences in a bilingual corpus. Compared with the manually constructed Chinese FrameNet, our approach results a higher coverage in terms of LUs and sample sentences. Furthermore, filtering strategies are explored to reduce the noise from the automatic generated data. Human verification confirms the quality of the outcomes. Our methodology can be easily adapted to generate the FrameNet-style dataset for other languages.

In the rest of this paper, we first introduce the linguistic resources that support our construction in Section 2. Section 3 presents our methodology and discusses the filtering strategies that are used for quality improvement. Section 4 demonstrates an application of our resource that would improve the annotation process in terms of annotation time consumption. Finally, we conclude our contributions and discuss future work in Section 5.

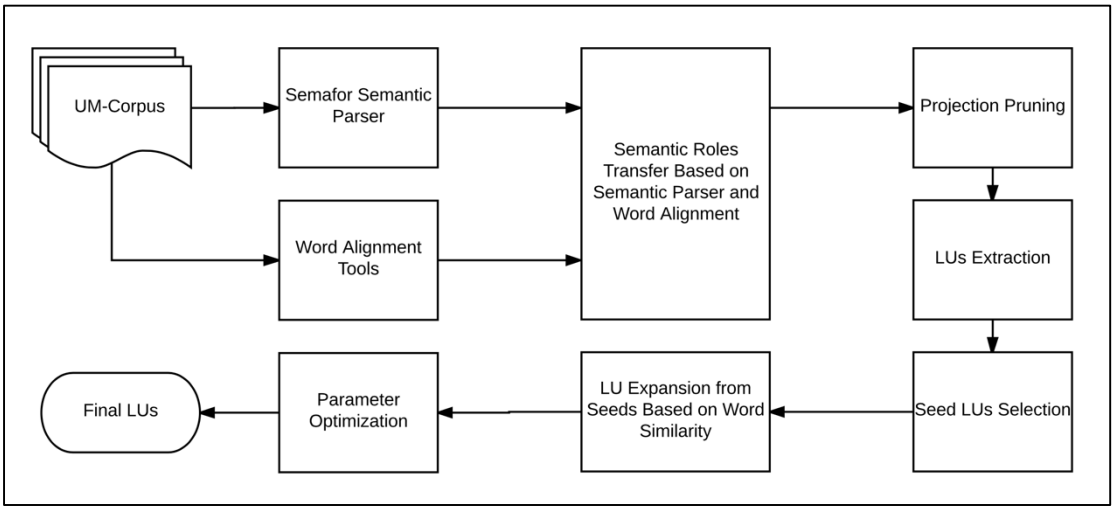

Figure 1: Overview of our approach to Chinese FrameNet construction.

2. Resources

We construct Chinese FrameNet based on UM-Corpus (Tian et al., 2014), which is a large-scale, balanced English-Chinese corpus consisting of 2.2 million parallel sentences from eight genres in a reasonable proportion, including News, Spoken, Laws, Thesis, Education, Science, Subtitle, and Microblog. They are parsed and extracted from online journals (national and international), official websites, online language learning resources (e.g. online dictionary and translation portals), TED, and Microblogs. Tian et al. (2014) apply some well-designed algorithms and tools to speed up the building process, such as document alignment, sentence boundary detection, and sentence alignment. The constructed corpus is manually verified to ensure the quality.

3. Methodology

Figure 1 depicts the overview of our approach to automatic Chinese FrameNet construction. Based on a sentence-aligned bilingual corpus, our basic idea is to transfer the machine-annotated frame information from English sentences to their Chinese counterparts. In Section 3.1, the English semantic parser, SEMAFOR, is performed to label the FEs on English sentences. In Section 3.2, we project the FEs from the English part to the Chinese part, based on the word alignment generated by TsinghuaAligner. The invalid projections are truncated by the strategy described in Section 3.3. In Section 3.4, the candidate LUs are extracted from the machine-annotated Chinese samples. From the candidate LUs in Section 3.4, we further select the frequent ones as the seeds. In Section 3.5, an expansion algorithm is proposed to augment LUs based on the seeds.

3.1 English Semantic Frame Labeling

We label the English part of UM-Corpus with semantic frames for the later step, projection. Here, we utilize SEMAFOR (Das et al., 2010), a state-of-the-art semantic frame parser for English based on the FrameNet ontology. The tool finds the words that are likely to evoke frames, and then labels frame elements for each frame using a log-liner model.

3.2 Projection of Bilingual Frame Elements

Word alignment tools align words in a sentence in the source language to the corresponding words in the sentence in the target language. In our case, we regard English as the source language and Chinese as the target language since we will utilize the alignment information as our basis of finding the projection of semantic relations from English to Chinese between bilingual sentences. In this paper, we employ TsinghuaAligner (Liu and Sun, 2015), which takes the translation probabilities derived from GIZA++ (Och and Ney, 2003) as the central feature in the word alignment process.

The alignment procedure consists of three major steps. Figures 2, 3, and 4 show examples for the tasks respectively. First, we label 2.2M English sentences in the bilingual UM-Corpus with frame semantics using SEMAFOR. As shown in Figure 2, SEMAFOR generates 14M predicate-argument structures from these 2.2M sentences. Figure 3 shows the second step, where TsinghuaAligner is performed to derive English-Chinese word pairs as our basis for the step. Finally, as illustrated in Figure 4, we utilize both the parsed result and the alignment information gathered from SEMAFOR and TsinghuaAligner to produce Chinese FrameNet-style annotations.

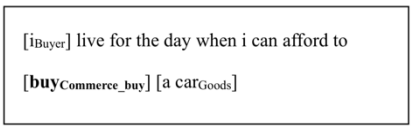

Figure 2: Applying the SEMAFOR parser to UM-Corpus.

Figure 3: Applying TsinghuaAligner from English to Chinese.

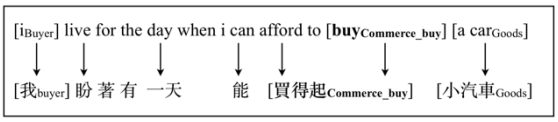

Figure 4: Mapping semantic roles from English to Chinese.

3.3 Pruning off Invalid Projections

From the alignment of bilingual frame elements, we have produced our preliminary results. However, not all frame structures produced in Section 3.1 are successfully projected due to imperfect English-Chinese alignments. Typical projection errors include the following two types:

(1) Missing target or frame elements.
The alignment tools we used do not guarantee full alignment coverage for each word in a sentence. Therefore, some target or frame elements are not projected from English to Chinese if no alignment information is provided.

(2) Incorrect frame elements projection.
Unlike target word in a frame, frame elements may consist of a group of words (i.e., phrase). However, it is possible that some of the words in the group cannot be aligned due to missing alignment information.

To alleviate the errors resulted from word alignments in (1) and ensure the quality of the projected sentences, we set a constraint on the number of frame elements in Chinese projected frame structures, which should be identical to that of English annotations. Based on this constraint, we discard the Chinese projections with missing frame elements. About 34% of the frame structures produced by SEMAFOR are removed in this process due to the aforementioned alignment errors. As a result, we obtain a dataset that consists of nearly 9M Chinese FrameNet-style annotations.

3.4 Lexical Unit Extraction

After aligning and pruning procedures, we have 9M Chinese FrameNet-style annotations as the basis for analysis. There are 779 unique frames in the 9M annotated instances. Compared with the current English FrameNet (FrameNet 1.7), which consists of 1,221 frames, the shortage results from SEMAFOR semantic parser, whose model was trained on 779 frames in FrameNet 1.3. We then locate the target of each frame from the 9M Chinese annotations, and produce 1M candidate LUs. The candidates may contain noise, so we refine them with the following steps.

3.5 Filtering of Lexical Units

We attempt to select reliable LUs from the candidates as the seeds for expansion. Here, we propose a statistical-based approach where the most frequent *N* LUs in a frame are regarded as seed LUs for the frame. We set *N* to be sufficiently small (*N*=10 in our setting) to ensure the quality. The setting produced 7,401 seed LUs, where 93.5% of them are considered valid after human verification. Table 1 shows some examples. Many errors are due to wrong word sense disambiguation in SEMAFOR. For instance, the meaning of the word "lead" should be the winning position during a race, while it is mis-resolved as the heavy, soft, dark grey metal by SEMAFOR. Another issue we noticed is that some word sense in English could not be directly mapped to Chinese. For example, the word 上海 "Shanghai" in English may mean to force someone to do something or go somewhere, while this word only carries the location sense in Chinese. By removing such errors, we could proceed to expand more LUs from the seeds without introducing too much noise.

Frame Name	LU seeds
Being_in_category	看作 (regard as), 相當於 (equivalent to), 等於 (equal to), 等價 (equivalent to), 無異於 (tantamount to), 等同於 (same as)
Food	食物 (food), 食品 (grocery), 蘋果 (apple), 咖啡 (coffee), 魚 (fish), 水果 (fruit), 脂肪(fat), 蔬菜(vegetable), 麵包 (bread)
Inclination	傾向 (tendency), 趨勢 (trend), 發展趨勢 (development trend), 傾斜 (tilt), 斜 (oblique), 變化趨勢 (trend), 趨向 (tend), 方向 (direction), 走向 (direction)
Planting	播下 (sow), 播種 (sow), 播種面積 (seeded area), 種子(seed), 撒(sprinkle)
Traversing	通過 (pass),傳遞 (transfer), 經過 (pass by), 上升 (rise), 交叉 (cross), 過去 (pass through), 穿越 (pass through), 下降 (decline), 跳 (jump), 遍歷 (traverse)
Visiting	客人 (guest), 訪客 (visitor), 嘉賓 (guest), 來賓 (guest), 遊客 (guest), 賓館 (hotel), 賓客 (guest), 招待所 (guest house)

Table 1: Examples of verified LU seeds

3.6 Expansion of Lexical Units

The seed LUs for each frame provide a great basis for the expansion. We expand the LUs based on word embedding. A CBOW word model is trained on UM-Corpus. For each seed LU in a certain frame, we could find a set of similar words based on cosine similarity. To ensure the quality, we set a threshold *t*, which is the minimum cosine similarity that a word could be added to the expanded LUs. Different threshold *t* will produce different LU size with different quality. In order to find an optimal setting, we sample 100 LUs from 100 frames for various threshold *t* as an index of the performance. Figure 5 shows the relationship between the LU size and the quality. The union of the all the sets within each frame forms the final LUs. The growth of the

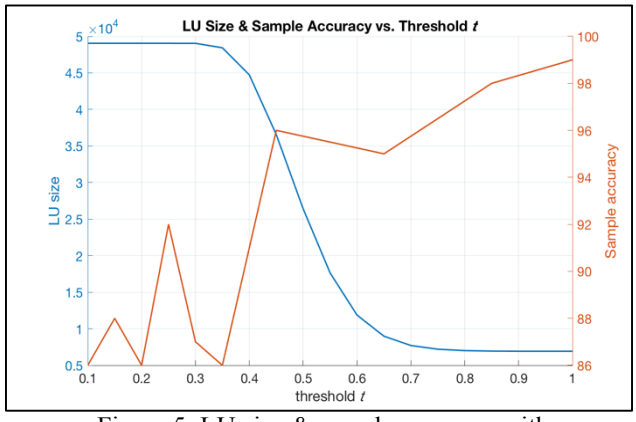

Figure 5: LU size & sample accuracy with different threshold *t*.

union is steady and will not expand too much since each seed in a frame conveys the similar meaning, resulting a union of many highly overlapping sets. We select t=0.45 as our final setting, which achieves a sample accuracy of 96% and yields 36k LUs. Table 2 shows the result of the expansion for the frame **Commerce_buy**.

LUs for frame **商業購買/Commerce_buy**	
CFN	購 (buy)/v; 購買 (purchase)/v; 購物 (shop)/v; 買(buy)/v; 租 (rent)/v; 租賃 (rent)/v; 租用 (lease)/v;
Our Resource	訂購 (order)/v; 參股 (share)/v; 買下 (buy in)/v; 收購 (buy in)/v; 買進 (buy in)/v; 買來 (buy in)/v; 入股 (share)/v; 買到 (buy)/v; 花錢買 (spend money on)/v; 買下來 (buy)/v; 買不到(unable to buy)/v; 支付 (transfer)/v; 交易 (transaction)/n; 並購 (merge)/v; 選購 (purchase)/v; 買不起 (cannot afford)/v; 買得起 (afford)/v; 採購 (purchase)/v; 購買者 (buyer)/n; 買過 (bought)/v; 買進賣出 (buy and sell)/v; 購入 (buy in)/v; 搶購 (rush to buy)/v; 購買 (purchase)/v; 買斷 (buyout)/v; 買賣 (trade)/v; 買家 (buyer)/n; 買 (buy)/v; 購進(buy in)/v; 買入(buy in)/v; 購得/v; 買個 (buy one)/v; 買些 (buy some)/v; 投資(invest)/v; 購置(purchase)/v;

Table 2: The expansion results for frame **Commerce_buy.**

Figure 6: Annotation interface that recommends candidate frames from a target word.

4. Application

In one subtask of our corpus annotation campaign, annotators are requested to provide Chinese FrameNet annotation on daily events from diaries and tweets. Annotators need to select a few target words (predicate of events) from a text first, and then annotate each target words with the correspondent FrameNet annotations. To annotate a frame, one must select which frame a target word could evoke, and realize its frame elements. Our constructed LUs are particularly useful for recommending frames for a target word (i.e, given a target word, list its possible frames by querying if the word could be found for some frames as LUs in our LU resource). Figure 6 depicts part of the annotation interface that recommends candidate frames given a target word. With low coverage LUs, the system often fails to find appropriate frame recommendations. In such case, annotators would need to spend considerable amount of time just to select a proper frame from a list of more than a thousand, which may severely slow the whole annotation process. By applying our resource to the annotation system, we could help alleviate the difficulty of selecting proper frame resulted from lack of LU resource.

We hired a few highly trained linguists to fulfill our corpus annotation. At the time of writing we have 36,029 target words marked as event predicates. 30,976 out of all target words have its FrameNet annotation. Table 3 shows the annotation statistics for different source types. We achieve overall 86% success rate to provide frame recommendations that are acceptable by annotators. The success rates in terms of source types are 95% and 80% for diaries and tweets respectively. Since our LU resource is constructed from UM-Corpus, which consists of contents that are relatively formal, the result also shows much higher success rate for diaries compared with tweets.

	# Texts	# Predicate	# Frames	%Success
Diaries	4,688	13,451	12,793	0.95
Tweets	26,818	22,578	18,183	0.80
Total	31,506	36,029	30,976	0.86

Table 3: Annotation statistics for different source types.

5. Conclusions and Future Work

This paper presents the development process and the current status of the construction for a large scale FrameNet resource in Chinese. Based on a bilingual corpus, we construct a more applicable resource that has lexical units with higher coverage that could help improve the annotation process in terms of efficiency. With more annotated dataset available in the future, we could develop a more robust automatic semantic role labeling tool for Chinese.

6. Acknowledgements

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-105-2221-E-002 -154 -MY3, MOST-106-2923-E-002 -012 -MY3 and MOST-107-2634-F-002-011-, and Academia Sinica under grant AS-107-TP-A05.

7. References

Das, D., Schneider, N., Chen, D., & Smith, N. A. (2010, June). Probabilistic frame-semantic parsing. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 948-956). Association for Computational Linguistics.

Fillmore, C. J. and Baker, C. (2009). A frames approach to semantic analysis. In B. Heine & H. Narrog (Eds), *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press, pp. 313–339.

Fillmore, C. J., Baker, C. F., and Sato, H. (2002). The FrameNet Database and Software Tools. In Proceedings of the 2002 International Conference on Language Resources and Evaluation (LREC 2002), pages 1157–1160. European Language Resource Association (ELRA).

Fillmore, C. J., (1982). Frame Semantics. In *Linguistics in the Morning Calm*. Seoul, South Korea: Hanshin Publishing Co., pp. 111–137.

Hermann, K. M., Das, D., Weston, J., and Ganchev, K. (2014). Semantic Frame Identification with Distributed Word Representations. In Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics (ACL 2014), pages 1448–1458, Baltimore, Maryland, USA. Association for Computational Linguistics (ACL).

Kim, J.-U., Hahm, Y., and Choi, K.-S. (2016). Korean FrameNet Expansion Based on Projection of Japanese FrameNet. In Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016): System Demonstrations, pages 175–179, Osaka, Japan, December.

Kingsbury, P. and Palmer, M. (2002). From TreeBank to PropBank. In Proceedings of the 2002 International Conference on Language Resources and Evaluation (LREC 2002), pages 1989–1993. European Language Resource Association (ELRA).

Li, R., Wu, J., Wang, Z., and Chai, Q. (2015). Implicit Role Linking on Chinese Discourse: Exploiting Explicit Roles and Frame-to-Frame Relations. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL 2015), pages 1263–1271, Beijing, China, July. Association for Computational Linguistics (ACL).

Liu, Y. and Sun, M. (2015). Contrastive Unsupervised Word Alignment with Non-Local Features. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015), pages 2295–2301. AAAI Press.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics,* 29(1):19–51.

Ohara, K. H., Fujii, S., Saito, H., Ishizaki, S., Ohori, T., and Suzuki, R. (2003). The Japanese Framenet Project: A Preliminary Report. In Proceedings of Pacific Association for Computational Linguistics, pages 249–254.

Park, J., Nam, S., Kim, Y., Hahm, Y., Hwang, D., and Choi, K.-S. (2014). Frame-Semantic Web: A Case Study for Korean. In Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272, pages 257–260. CEUR-WS. org.

Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., Li, S., Wang, Y., and Lu, Y. (2014). UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland.

Tonelli, S. and Pianta, E. (2008). Frame Information Transfer from English to Italian. In Proceedings of the 2008 International Conference on Language Resources and Evaluation (LREC 2008), pages 2252–2256. European Language Resource Association (ELRA).

You, L. and Liu, K. (2005). Building Chinese FrameNet Database. In Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE 2005).