English Event Detection With Translated Language Features

Sam Wei
School of IT
University of Sydney
Sydney, Australia
swei4829@uni.sydney.edu.au

Igor Korostil
TEG Analytics
Sydney, Australia

Joel Nothman
Sydney Informatics Hub
University of Sydney
Sydney, Australia

Ben Hachey
School of IT
University of Sydney
Sydney, Australia
{eeghor,joel.nothman,ben.hachey}@gmail.com

Abstract

We propose novel radical features from automatic translation for event extraction. Event detection is a complex language processing task for which it is expensive to collect training data, making generalisation challenging. We derive meaningful subword features from automatic translations into target language. Results suggest this method is particularly useful when using languages with writing systems that facilitate easy decomposition into subword features, e.g., logograms and Cangjie. The best result combines logogram features from Chinese and Japanese with syllable features from Korean, providing an additional 3.0 points f-score when added to state-of-the-art generalisation features on the TAC KBP 2015 Event Nugget task.

1 Introduction

Event trigger detection is the task of identifying the mention that predicates the occurrence of an event and assigning it an event type (e.g., attack). Typical training data for event trigger detection includes fewer than 200 annotated documents (Ellis et al., 2015). Yet systems attempt to identify many event types (e.g., 38 for the data used here), making data sparsity a particular challenge (Ji, 2009; Zhu et al., 2014).

Existing approaches use two main strategies for handling data sparsity. One strategy is to use lexical databases. Lexical databases have become a standard feature set for event detection. They make it easy to include synonyms and word-class information through hypernym relations. However, they require substantial human effort to build and can have low coverage. Another approach is to induce word-class information through clustering. Here cluster co-membership can be used to find synonyms and cluster identifiers provide abstracted word-class information.

We propose novel semantic features for English event detection derived from automatic translations into thirteen languages. In particular, we explore the use of Cangjie[1] radicals in Chinese and Japanese. Where characters represent concepts, they have often been composed of smaller pictographic units, called radicals. For example: 明(bright) is composed of two radicals 日,月(sun, moon) with corresponding Latin letter sequence "AB". While this composition is often not productive, we hypothesise that the recurrence of some radicals among related concepts' logograms may be exploited to identify semantic affinity.

Results suggest that (1) translated language features are especially useful if the target language has a writing system facilitating easy decomposition into useful subword features; (2) logograms (e.g., Chinese, Japanese), radicals (e.g., Chinese, Japanese) and syllables (e.g., Japanese, Korean) prove beneficial and complementary; and (3) Chinese characters are particularly useful, comparable to WordNet. Adding the best translated language features to the final system improves F1 by 3.0 points over a state-of-the-art feature set on the TAC KBP 2015 nugget type detection task.

2 Background

Multilingual resources have been successfully applied to various NLP tasks such as named entity recognition (Klementiev and Roth, 2006), paraphrasing (Bannard and Callison-Burch, 2005), sentiment analysis (Wan, 2008), and word sense disambiguation (Lefever and Hoste, 2010).

[1] https://en.wikipedia.org/wiki/Cangjie_input_method

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 293–298
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-2046

Ji (2009) reports significantly improved event trigger extraction via cross-lingual clusters of English translations to Chinese trigger words over large corpora. At runtime, these are used to replace low-confidence event triggers with other high-confidence predicates from the same cluster. We describe an approach leveraging cross-lingual information not only from words, but also at the level of characters and radicals. Like Zhu et al. (2014), we use Google Translate and build bilingual feature vectors from the translations as well as original English sentences. While they address event trigger type classification only, we address both trigger detection and classification. We use new translated language features and evaluate with a range of languages.

Li et al. (2012) show that monolingual Chinese event trigger extraction benefits from using compositional semantics inferred from Chinese characters. We use similar Chinese character information as features for English event trigger detection also using maximum entropy modelling. Furthermore, we introduce new radical features that take advantage of semantic compositionality of Chinese characters.

2.1 Task

We address the event nugget detection task from the Text Analysis Conference Knowledge Base Population (TAC KBP) 2015 shared task (Mitamura and Hovy, 2015), which includes trigger detection and classification. An event trigger is the smallest extent of text (usually a word or short continuous phrase) that predicates the occurrence of an event (LDC, 2015). The task defines 9 event types and and 38 subtypes. Like most task participants, we formulate event trigger detection as a token-level classification task. We use a maximum entropy classifier here, with IOB encoding (Sang and Veenstra, 1999) to represent multi-word mentions.

For comparison, we implement the baseline and lexical generalisation features from Hong et al. (2015). This was the best-performing system in the TAC 2015 nugget type detection task, with an F1 of 58.3. We do not replicate their semi-supervised techniques here as we want to isolate the comparison of translated language features to other generalisation features. Since translated language features leverage off-the-shelf automatic translation, we believe the results here will generalise to semi-supervised learning as well.

Baseline Features (BASE) Our baseline system uses standard surface features used for event extraction. Features of the current token include the full word token as it appears in the sentence, its lemma, its part of speech (POS), its entity type, and a feature that indicates whether the first character of the token is capitalised. Context features are computed for a window of one token on either side of the current token. They include lemma bigrams, POS bigrams and entity type bigrams. Finally, grammatical features are computed based on a dependency parse of the sentence. These include dependency relation types for the governor and any dependents, conjoined relation type and lemma, conjoined relation type and POS, and conjoined relation type and entity type.

Lexical generalisation Features (LEX) We include three generalisation feature sets from the literature as a benchmark. The first lexical resource we use is Nomlex (Macleod et al., 1998) – a dictionary of nouns that are generated from another verb class, usually verbs. We also use Brown clusters trained on the Reuters corpus (Brown et al., 1992; Turian et al., 2010). Brown clusters group words into classes by performing a hierarchical clustering over distributional representations of the contexts in which they appear. Finally, we use WordNet (Miller, 1995) – a lexical database that includes synonym relations and semantic type-of/hypernym relationships. These relations have been used to extend feature sets beyond observed tokens which can help with identification of rare or unseen event triggers.

3 Approach

We use machine translation (MT) service to obtain translated text. The translation is done at sentence level. We cache the translation results on files to ensure the experiments are repeatable. Below are example sentences translated from English into Chinese and Spanish.

EN *The attack by insurgents happened yesterday.*
ZH 叛亂分子的襲擊發生在昨天。 (1)
ES *El ataque de los insurgentes pasó ayer.*

3.1 Translated Language Features (TRANS)

We generate three types of logogram features and use stem features for non-logogram languages.

Word features (word) Different words in English can be translated into the same word in another language. For example there are 201 unique

Chinese Character	Radical Symbol	Latin Radical	English Word
打	手一弓	QMN	hit
擊	十水手	JEQ	strike
投	手竹弓水	QHNE	throw
擲	手廿大中	QTKL	throw
折(磨)	手竹一中	QHML	torture
拆	手竹一卜	QHMY	demolish
拷(打)	手十大尸	QJKS	torture
割	十口中弓	JRLN	cut
刺	木月中弓	DBLN	stab

Table 1: Attack event triggers. The radical "手" (Q, hand) frequently appears in the attack event triggers. Radicals "中弓(刀)" (LN, knife) appear frequently when events are associated with actions that are performed with a knife

English trigger words for attack events and only 160 unique words in their Chinese translations. Therefore if an English trigger word is not in the training data, the model might still recognise the trigger if it has seen the Chinese translation before.

Logogram character features (char) Chinese and Japanese logograms are compositions of one or more characters defining their meanings. Therefore, different words representing the same event often contain similar characters. There are 195 unique Chinese characters for the attack event triggers in the corpus. The most frequently appearing characters are "擊" (strike, attack), "戰" (war, fight), "殺" (kill), "爭" (fight, dispute), and "炸" (bomb, explode).

Logogram Cangjie features (Cangjie) Chinese and Japanese characters can be further decomposed to smaller components called radicals. Certain radicals are more commonly found for a particular event type (Table 1). Cangjie is one of the methods to decompose Chinese characters. It was designed to use on computers with QWERTY keyboards so the radicals can be easily stored, indexed and searched by most computer systems. In addition to word and character features, we compute Cangjie features for logographic languages.

Stem features (stem) For many languages character and radical features cannot be generated. We generate stem features in addition to the word features where available. We use the NLTK Snowball stemmer for German, Spanish, Finnish, Hungarian, Dutch and Russian; and the NLTK ISRI stemmer for Arabic. By including a range of languages, we hope to separate the effect of syllabic from semantic components of logograms.

3.2 Translation Alignment

Translated language features require each English word to be aligned to one in the translated sentence. We use the translation service obtain all possible translations of a given English word, e.g.:

EN *attack*
ZH 進攻, 砰擊, 發作, 攻擊, 攻打, 掊擊, 抨擊, ... (2)
ES *acometida, ataque contra, agresión, ...*

If one of these is in the translated sentence, then an alignment is made. If not, then we use the most likely word translation (underlined above).

4 Experiments

We use the TAC KBP 2015 English event nugget data (Ellis et al., 2015) for the experiments. Development experiments use the training data (LDC2015E73) and the evaluation data (LDC2015R26) is held out for final results. The development corpus contains a total of 158 documents from two genres: 81 newswire documents and 77 discussion forum documents. We split this into 80% for training and 20% for development testing. We use Google Translate to obtain sentence and word translations into target languages and derive translated language features to help with the English task. Evaluation uses the official scorer from the shared task, where a trigger is counted as correct if both the trigger span and its event subtype are correctly identified.

Comparing languages First, we explore how translated language features perform across the thirteen languages. Figure 1 shows how much each target language improves BASE on development data. We include all word, stem, character and Cangjie features as available for each language. Chinese, Japanese and Korean stand out, with improvements as high as 19.17 points f-score due mostly to large increases in recall. These results suggest that languages with writing systems that facilitate easy decomposition into meaningful subword features are particularly useful.

Combining languages Next, we test whether system performance can be further improved using TRANS features from multiple languages. We add target languages one at a time in order of individual performance, and find that Traditional Chinese, Japanese and Korean to Simplified Chinese together improve F1 by 2.5 points. This combined feature set is used in the remaining analysis and experimental results.

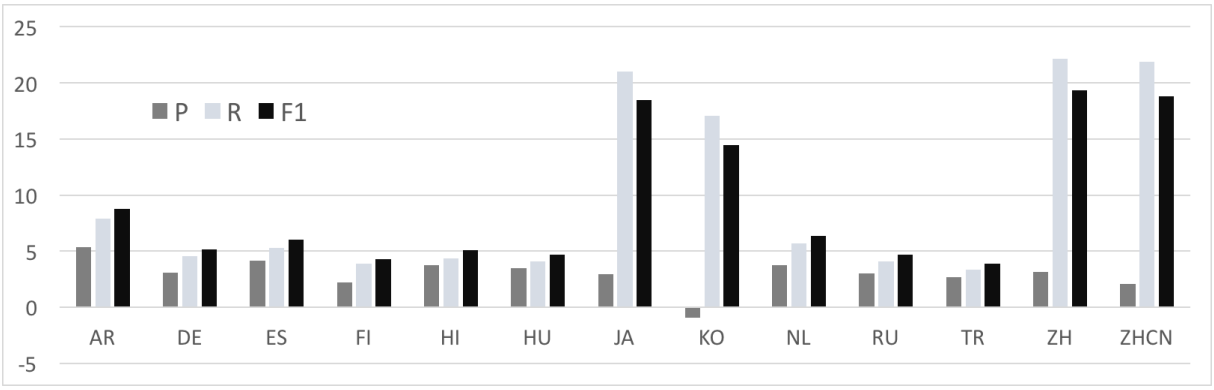

Figure 1: Effect of individual languages on development data, showing the difference in precision, recall and F1 compared to the BASE scores of 55.16, 20.62 and 30.02. AR:Arabic, DE:German, ES:Spanish, FI:Finnish, HI:Hindi, HU:Hungarian, JA:Japanese, KO:Korean, NL:Dutch, RU:Russian, TR:Turkish, ZH:Chinese (Simplified), ZHCN:Chinese (Traditional).

Error analysis We explore characteristic errors for BASE+LEX versus BASE+TRANS for the attack event on evaluation data. We randomly sample twenty instances where one is correct and the other is incorrect. Of six LEX FN errors, two are triggers not seen in the training data, e.g., 'wages' (*Transfer-Money*), and 'resignation' (*End-Position*). In other cases, there seem to be too few training instances, e.g., 'pardoning' (*Pardon*) only appears once in the training data. The TRANS FN error is due to a bad translation in which 'strike' (*Attack*) is a translated to the 'work stoppage' sense instead of the 'forceful hit' sense.

For both systems, most FP errors correspond to cases with challenging ambiguity. For instance, both systems label 'appeal' as *Justice.Appeal* event in two sentences where the word 'appeal' means 'ask for aid', instead of 'taking a court case to a higher court'. The translation was incorrect in this case. Similarly, 'report' appears six times in the training data as three different event types (*Broadcast*, *Correspondence*, *Move-Person*).

Long-tail generalisation Table 2 shows type-level results for BASE+LEX and BASE+TRANS compared to BASE alone. The generalisation feature sets outperform the baseline for all but three of the 38 event types. For *Pardon*, BASE obtains 97 F1 so there is little room for improvement. For *Execute*, LEX features have no effect while TRANS doubles BASE F1. *Contact* is the only type where generalisation features are harmful. Ignoring ties, BASE+TRANS performs best on more types (13) than BASE+LEX (11). TRANS appears to help more with long-tail entity types that have fewer training instances (e.g., *Bankruptcy*, *Appeal*, *Born*). Encouragingly, this analysis also suggests that LEX and TRANS can be complementary, with LEX doing particularly well on some types (e.g., *Trial-Hearing*, *Correspond*) and TRANS doing particularly well on others (e.g., *Transfer-Money*, *Release-Parole*).

Type	Trn	Tst	BA	LX	TR
Attack	547	253	29	**60**	58
Move-Person	390	127	15	**37**	33
Transfer-Money	366	185	18	35	**49**
Die	357	157	45	63	**66**
Broadcast	305	112	14	**20**	16
Contact	260	77	**29**	24	23
Transfer-Ownership	234	46	9	20	**34**
Meet	221	23	15	**44**	38
Pardon	221	18	**97**	**97**	95
Arrest-Jail	208	79	54	70	**71**
Convict	173	49	71	74	**81**
End-Position	130	79	18	51	**55**
Extradite	62	1	0	0	**100**
Execute	51	15	12	12	**24**
Release-Parole	45	28	0	87	**95**
Bankruptcy	30	3	0	50	**89**
Appeal	25	12	0	57	**92**
Born	13	6	0	22	**40**

Table 2: Comparing instance count in training (Trn) and test (Tst) to F1 for BASE (BA), LEX (LX) and TRANS (TR).

5 Final Results and Discussion

Table 3 contains final results on the held-out evaluation data. The final translated language feature set (TRANS) comprises word, character and Cangjie features from Traditional Chinese, Simplified Chinese, Japanese and Korean. TRANS features provide a large F1 improvement of 17.4 over the baseline (BASE), similar to the benchmark lexical generalisation features (LEX). They differ in precision-recall tradeoff, with higher recall but lower precision from TRANS. LEX and TRANS are complementary, giving F1 of 55.0.

System	P	R	F
BASE	60.4	24.1	34.4
BASE+LEX	66.8	42.6	52.0
BASE+TRANS	59.6	45.8	51.8
BASE+LEX+TRANS	67.9	46.2	**55.0**
TAC 2015 medians	61.7	40.7	48.8
TAC 2015 #1	75.2	47.7	**58.4**

Table 3: Final results comparing translated language features (TRANS) to benchmark lexical generalisation features (LEX). BASE+LEX is our implementation of the core Hong et al. classifier. TAC KPB 2015 #1 corresponds to reported results for Hong et al. including semi-supervised learning. TAC KPB 2015 shared task has 38 runs submitted from 14 teams.

This is 20.6 points higher than the baseline features alone, and improves both the precision of LEX and the recall of TRANS.

The main appeal of the approach here is that translated character and radical features are easy to obtain using off-the-shelf tools. This provides a simple technique to capture semantic information and leverage the word sense disambiguation encoded in translation models trained over very large datasets. Given the positive results here, we plan to explore translation and alignment strategies to improve precision. We also plan to quantify the effect of different translation systems and system change over time.

6 Conclusion

We described an event detection system leveraging features from off-the-shelf automatic translation to improve generalisation to new data. Chinese, Japanese and Korean prove especially useful as they provide natural decomposition into informative subword features, i.e., characters (Chinese and Japanese), radicals (Chinese and Japanese) and syllables (Korean). None of the nine other languages explored provide similar levels of natural decomposition and none provided additional benefit. The best system includes Chinese, Japanese and Korean character features. These translated language features improve f-score by 3 points on top of the English-only generalisation features from WordNet, Nomlex and Brown clusters.

Acknowledgments

We wish to thank Will Radford and the anonymous reviewers for their helpful feedback. This research is funded by the Capital Markets Co-operative Research Centre. Ben Hachey is the recipient of an Australian Research Council Discovery Early Career Researcher Award (DE120102900).

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Annual Meeting of the Association for Computational Linguistics*. pages 597–604. https://doi.org/10.3115/1219840.1219914.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18(4):467–479. http://www.aclweb.org/anthology/J/J92/J92-4003.pdf.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie Strassel. 2015. Overview of linguistic resources for the TAC KBP 2015 evaluations: Methodologies and results. In *Text Analysis Conference*.

Yu Hong, Di Lu, Dian Yu, Xiaoman Pan, Xiaobin Wang, Yadong Chen, Lifu Huang, and Heng Ji. 2015. RPI BLENDER TAC-KBP2015 system description. In *Text Analysis Conference*.

Heng Ji. 2009. Cross-lingual predicate cluster acquisition to improve bilingual event extraction by inductive learning. In *NAACL Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*. pages 27–35. http://www.aclweb.org/anthology/W09-1704.

Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics*. pages 817–824. https://doi.org/10.3115/1220175.1220278.

LDC. 2015. *Rich ERE Annotation Guidelines Overview*. Linguistic Data Consortium. Version 4.1. Accessed 14 November 2015 from `http://cairo.lti.cs.cmu.edu/kbp/2015/event/summary_rich_ere_v4.1.pdf`.

Els Lefever and Veronique Hoste. 2010. SemEval-2010 task 3: Cross-lingual word sense disambiguation. In *International Workshop on Semantic Evaluation*. pages 15–20. http://www.aclweb.org/anthology/S10-1003.

Peifeng Li, Guodong Zhou, Qiaoming Zhu, and Libin Hou. 2012. Employing compositional semantics and discourse consistency in chinese event extraction. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. pages 1006–1016. http://www.aclweb.org/anthology/D12-1092.

Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barrett, and Ruth Reeves. 1998. Nomlex: A lexicon of nominalizations. In *Euralex International Congress*. pages 187–193.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM* 38(11):39–41. https://doi.org/10.1145/219717.219748.

Teruko Mitamura and Eduard Hovy. 2015. *TAC KBP Event Detection and Coreference Tasks for English*. Version 1.0. Accessed 14 November 2015 from `http://cairo.lti.cs.cmu.edu/kbp/2015/event/Event_Mention_Detection_and_Coreference-2015-v1.1.pdf`.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *Conference of the European Chapter of the Association for Computational Linguistics*. pages 173–179. https://doi.org/10.3115/977035.977059.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Annual Meeting of the Association for Computational Linguistics*. pages 384–394. http://www.aclweb.org/anthology/P10-1040.

Xiaojun Wan. 2008. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pages 553–561. http://www.aclweb.org/anthology/D08-1058.

Zhu Zhu, Shoushan Li, Guodong Zhou, and Rui Xia. 2014. Bilingual event extraction: a case study on trigger type determination. In *Annual Meeting of the Association for Computational Linguistics*. pages 842–847. http://www.aclweb.org/anthology/P14-2136.