Parallel Chinese-English Entities, Relations and Events Corpora

Justin Mott, Zhiyi Song, Ann Bies, Stephanie Strassel

Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA 19104 USA
E-mail: {jmott,zhiyi,bies,strassel}@ldc.upenn .edu

Abstract

This paper introduces the parallel Chinese-English Entities, Relations and Events (ERE) corpora developed by Linguistic Data Consortium under the DARPA Deep Exploration and Filtering of Text (DEFT) Program. Original Chinese newswire and discussion forum documents are annotated for two versions of the ERE task. The texts are manually translated into English and then annotated for the same ERE tasks on the English translation, resulting in a rich parallel resource that has utility for performers within the DEFT program, for participants in NIST's Knowledge Base Population evaluations, and for cross-language projection research more generally.

Keywords: language resources, information extraction, entities, relations, events

1. Introduction

This paper introduces the parallel Chinese-English Entities, Relations and Events (ERE) corpora developed at the Linguistic Data Consortium (LDC) as a part of DARPA's Deep Exploration and Filtering of Text (DEFT) program. The DEFT program seeks to improve state-of-the-art capabilities in automated deep natural language processing, with a particular focus on technologies dealing with inference, causal relationships, and anomaly detection across several languages (DARPA 2012). Given the large number and variety of approaches to algorithm development within DEFT, we set out to define an annotation task that would be supportive of multiple research directions and technology evaluations, and that would provide a useful foundation for follow-on DEFT annotation tasks like entailment, inference and belief/sentiment.

The resulting Entities, Relations and Events annotation task has evolved over the course of the DEFT program, from a fairly lightweight treatment of entities, relations and events in text, to a richer representation of phenomena of interest to the program (Song et al. 2015). ERE corpora are used by DEFT performers as a general resource, and also serve as training data for several tracks within the Text Analysis Conference Knowledge Base Population (TAC KBP) evaluation series, which is open to non-DEFT participants conducted by the National Institute of Standards and Technology (NIST). TAC KBP aims to develop and evaluate technologies for building and populating knowledge bases from unstructured texts (NIST 2015). The ERE corpora provide a training resource for component evaluation tasks such as Entity Detection and Linking and Event Argument Linking. In keeping with the overarching goals of the DEFT Program, TAC KBP increasingly focuses on extracting information from multilingual resources (Ji 2010; NIST 2015), and the parallel ERE data described in this paper are particularly useful in this context.

The parallel ERE corpora have relevance beyond the specific objectives of DEFT and TAC KBP. Cross-language projection is an important means to bootstrap the transfer of annotation across multiple languages, and has been applied to many annotation schemes, both with bitext (Yarowsky et al., 2001; Wang and Manning, 2014; Zitouni, 2008; Ehrmann et al., 2011) and without bitext (Zirikly 2014). The creation of independently annotated parallel data sets, such as the ones described here, can serve as a standard by which to evaluate the efficacy of a transfer approach.

2. ERE Annotation Overview

The ERE annotation schema is derived from earlier related efforts like Automated Content Extraction (ACE) (Doddington et al., 2004; LDC, 2005; Walker et al., 2006). As in ACE, ERE exhaustively labels entities, relations and events along with their attributes according to a specified taxonomy. ERE annotation has been produced in two stages: Light ERE and Rich ERE.

2.2 Light ERE Annotation

Light ERE is designed to be a streamlined version of ACE to allow rapid annotation over multiple languages (Aguilar et al. 2014). For entities, only specific entities are annotated in Light ERE. Entities are assigned one of the following types: person (PER), organization (ORG), geopolitical entity (GPE), location (LOC) and title (TTL). Mentions are classed according to mention level: named (NAM), nominal (NOM) or pronominal (PRO). Entity mentions are coreferenced with one another as appropriate. Unlike ACE, the heads of nominal mentions are not explicitly marked.

Only asserted relations between entities are labeled; hypothetical, future and negated relations go unlabeled. The relation ontology for Light ERE consists of 4 types and 10 subtypes.

For events, only positive, asserted events are captured. Events are required to be bound to an anchor in the text (a "trigger word") and are also required to have one or more arguments present in the text. As with entities, event mentions are clustered together when they are coreferential.

2.2 Rich ERE Annotation

Rich ERE entity annotation expands many areas of Light ERE annotation, starting with a general increase in the number of items that can be tagged (Song et al., 2015). For entities, Rich ERE captures underspecified and generic entities, in addition to specific, and labels them along a specific/non-specific axis. Rich ERE also separates the Light ERE Location entity type into Facility (FAC) as well as Location types, with Facility being defined as human-made. As in ACE, the heads of nominal mentions are explicitly marked.

Argument Fillers, similar to Values in ACE, and are added to Rich ERE and serve as event and relation arguments that are not otherwise labeled as entities. For example, the event Justice.Sentence takes arguments for the crime committed and the sentence imposed, neither of which are annotated as entities. The use of Argument Fillers allows the annotation to capture those arguments. Titles have been reclassified from an entity type in Light ERE to an Argument Filler in Rich ERE, which can then take part in relevant Relations, such as Social.Role, and Events, such as Personnel.StartPosition.

Rich ERE relations have an expanded ontology to better align with TAC KBP Slot Filling. Rich ERE has a total of 5 types and 20 subtypes. Hypothetical, future, conditional generic relations are annotated in addition to actual attested relations. Such relations are tagged with the Realis:Other attribute.

Rich ERE event annotation includes increased taggability in several areas: an expanded event ontology, with 9 types and 38 subtypes. Rich ERE includes the addition of generic and other (irrealis), such as future, conditional, hypothetical and negated, event mentions and the marking of irrealis state for arguments when the link between the event and argument is negated, hypothetical, etc. Event mentions no longer require the presence of an argument to be taggable. Contact and Transaction events are augmented with additional attributes.

One further extension of Rich ERE is the inclusion of "double tagging", i.e., the same event mention trigger span may be tagged more than once for different event types/subtypes when the trigger instantiates different event types or subtypes. It also allows the same type/subtype of event to be tagged more than once in certain coordinated structures. For example, the trigger "murder" in the example below is the trigger for two Life-Die events, one with the victim "George Besse" and the other with "Rene Audran", and two Conflict-Attack events, one with the time argument of 1986 and one with the time argument of 1985.

> *Cipriani was sentenced to life in prison for the murder of Renault chief George Besse in 1986 and the head of government arms sales Rene Audran a year earlier.*

Rich ERE has replaced strict event coreference with the concept of *Event Hopper*, which is a more inclusive, less strict notion of event coreference. Event hoppers contain mentions of events that are intuitively coreferential to the annotator even if they do not meet the earlier strict event identity requirement. This allows for event mentions to be grouped together even when the event arguments and/or temporal and location properties are represented at different levels of granularity in the text. For example, an event hopper for an Attack event could contain event mentions with the location arguments *Iraq*, *Baghdad* and *the Green Zone,* despite their differing levels of granularity.

Table 1 shows a parallel pair of Chinese-English sentences annotated for Rich ERE. Note for the sake of space and clarity, entity mention level and specificity are not shown.

Sentence	德国总理默克尔到中国来为什么？	What is German Chancellor Merkel coming to China to discuss?
Entities	**Entity-C1 (GPE)** 德国 **Entity- C2 (PER)** 德国总理 默克尔 **Entity-C3 (GPE)** 中国	**Entity-E1 (GPE)** German **Entity-E2 (PER)** Merkel **Entity-E3 (GPE)** China
Relations	**Relation-C1** Trigger: 总理 Realis: Asserted Type: Org-Affiliation Subtype: Leadership Argument 1: 默克尔 Argument 2: 德国 **Relation-C2** Trigger: 到	**Relation-E1** Trigger: Chancellor Realis: Asserted Type: Org-Affiliation Subtype: Leadership Argument 1: Merkel Argument 2: German **Relation-E2** Trigger: coming

	Realis: Other Type: Physical Subtype: Located-Near Argument 1: 默克尔 Argument 2: 中国	Realis: Other Type: Physical Subtype: Located-Near Argument 1: Merkel Argument 2: China
Event Hoppers	**Event-mention-C1** Trigger: 到 Realis: Other Type: Movement Subtype: TransportPerson Person: 默克尔 Destination: 中国	**Event-mention-E1** Trigger: coming Realis: Other Type: Movement Subtype: TransportPerson Person: Merkel Destination: China

Table 1: Example of Rich ERE parallel annotation

3. Parallel Chinese-English ERE

The parallel Chinese-English data set consists of 171 Chinese source files paired with the corresponding English translations. The data consists of approximately 100,000 words of Chinese translated into English under DARPA's BOLT program (Garland et al., 2014). The translation was high-quality manual translation and was also aligned at sentence level. Because the data was intended to support machine translation system development, the translations were specified to prioritize meaning fidelity over fluency. That is, translators were instructed to neither add nor remove any information content from the source sentence when creating the translation. This data set was in the internet discussion forum genre. ERE annotation was performed on each side independently. The overall data volume of the parallel ERE annotation is shown in Table 2.

	Chinese	English
Files	171	171
Characters	127,458	--
Words	--	101,191

Table 2: ERE parallel data volume

3.1 Parallel Chinese-English Light ERE

Light ERE annotation has been completed on both the Chinese and English sides. The data was produced using LDC's standard ERE pipeline of first pass annotation, followed by second pass annotation by experienced annotators with a subsequent corpus-wide quality control (QC) check. Kulick et al. (2014) provides a discussion of inter-annotator (IAA) procedures in the context of Light ERE.

3.1.1 Cross-Lingual QC for Light ERE

After a standard corpus-wide QC was performed independently on both the Chinese and English sides of the corpus, an additional corpus-wide cross-lingual QC was performed. This consisted of generating statistics for each pair of parallel files for the various layers of annotation: number of entity mentions broken down by mention level and type; number of entities broken down by type; number of relations broken down by type/subtype. etc.

Pairs of parallel files were flagged for further review when one side of the corpus contained a significantly higher or lower number of annotations. Senior Chinese annotators reviewed the flagged pairs. Errors in the Chinese side were corrected; potential errors in the English side were flagged and then reviewed by a native English speaking lead annotator and corrected as needed. In all, 25 of the 171 pairs of files were reviewed and corrected in this way. The corrections made in this step typically involved errors of omission: where one side created an annotation whereas the other side did not.

In addition to the identification of inconsistencies that would otherwise not have been corrected, this step shed some light on the types of differences between the languages that led to imperfect matching in the annotations as well as some effects of translation (see Section 4).

3.2 Parallel Chinese-English Rich ERE

Rich ERE annotation took as input the annotation created in Light ERE, with annotations then manually added to meet the expanded scope of Rich ERE. This again followed the same annotation pipeline of first pass, second pass and standard QC. Due to time constraints, cross-lingual QC was not performed again at the Rich ERE annotation stage.

4. Comparing Chinese and English ERE Annotation

Despite the extra QC step discussed above, there were still discrepancies between the numbers of tagged items on the two sides of the corpus. The patterns provided below are not meant to exhaustively explain systemic differences in the languages that led to differences in annotation. Rather, they are patterns that recurred in the QC pass frequently enough to merit mention by annotators. No attempt is made here at a robust categorization of differences. Also note that the translations in the following examples are taken from the translated section of the corpus as described in Section 3. Some of differences in the number of annotations are due to linguistic differences between the languages. It is also the case that many of the mismatches were due to incomplete and/or insufficient translations. Overall, many of the examples uncovered display differences of granularity/specificity between the source and the translation.

As a simple example, because the phenomenon of pro-drop is more common in Chinese than in English there are fewer pronominal mentions in the Chinese data, as in the example below. The Chinese side has only two tagged pronominal mentions (***bolded***), whereas there are three in the English.

> *严重支持 lz 的观点，但是**我们**的政府和军人就是这样的不争气，**谁**也没有办法*
>
> ***I** solemnly support the opinion of lz, but **our** government and troops are just this disappointing, and there is nothing **we** can do*

Table 3 below shows the number of entity mention types in both Light ERE and Rich ERE. Overall in both Light and Rich ERE data, the Chinese side has many fewer pronominal mentions than the English side.

Translation effects not surprisingly led to differences in the number of items annotated. For example, when the English side encountered the entity translated as *South China Sea* it tagged a physical relation between *South China Sea* and *China*. However, the name of the same entity in Chinese is literally *South Sea (南海 nánhǎi)*. Because there is no entity mention for *China*, no relation could be annotated.

	Chinese Light	English Light	Chinese Rich	English Rich
NAM	6,570	5,063	6,675	5,211
NOM	3,558	2,669	5,106	5,853
PRO	1,899	3,345	2,321	4,991

Table 3: Light and Rich ERE Entity Mention Types

In some cases the specificity of the translation compared to the source, even when meaning fidelity was preserved, led to annotation differences. Below, in the Light ERE framework, a title must resolve to a specific position to be taggable. Therefore, *a chief* on the English side was considered too generic to support a title mention, whereas on the Chinese side *处长 chùchàng* always denotes a specific position and so was tagged.

> *民进内蒙古委员会社会服处**处长**石某*
>
> ***A chief** of the United Front Work Department of Inner Mongolia*

Imperfect matches in the English translations for the sake of grammaticality contributed to some mismatches between the two sides. For example, in the following the English translation has provided the text *his victim* to render it grammatical**,** which is not present in the Chinese. This results in not just an entity mismatch between the sides; because *victim* can serve as a trigger for a Conflict.Attack event, there is a mismatch in the number of annotated events as well (word-by-word gloss provided here).

> *马杀人的时候，是以平视，甚至仰视的角度，而药......是俯视*
>
> *Ma kill people DE when/time, is from “look at on the same level"or even “look up” DE angle, but Yao...... is look down.*
>
> *When Ma committed murder, he treated his victims as equal and even with admiration, but Yao... he looked down on **his victim***

Table 4 shows the difference in the annotation counts for entities, relations and events in Light and Rich ERE. For both Chinese and English, there are many more entities, relations and events annotated in Rich ERE due to the added taggability and enlarged taxonomy. For Light ERE, there is more annotation on the Chinese side than on the English side at all levels. However, for Rich ERE, there is more annotation on the English side than on the Chinese side at all levels, except at the entity level, in which Chinese has slightly more entities annotated than English. The Rich ERE data has roughly twice as many annotated events on the English side as on the Chinese side. A more systematic study of the differences, especially the divergence in the number of events annotated is planned for future work.

	Chinese Light	English Light	Chinese Rich	English Rich
Entity Mentions	12,206	11,231	14,102	16,055
Entities	4,984	3,523	5,974	5,873
Fillers	--	--	607	906
Relations	1,595	1,189	1,946	2,092
Event Mentions	481	369	1,491	2,933
Event Coref	391	308	1,138	2,285

Table 4: Light and Rich ERE tagged items

5. Annotation Decisions for Translation Artifacts

As mentioned in Section 3 above, ERE data used in this corpus was collected and translated under DARPA's BOLT program, which focused on machine translation. The emphasis on maximizing meaning fidelity over fluency in the translations resulted in some features that ERE had to accommodate by developing specific new annotation policies.

5.1 Alternate Translations

The English translations of this data have 274 instances where both fluent and literal translations of some Chinese expressions are present (Bies et al., 2014). The inclusion of the alternates was intended to assist machine translation system development. Below, the Chinese phrase *老毛子 lǎo máozi* is rendered with a fluent and literal translation:

> *为了目前我们既得利益，**老毛子**的事可以先不考虑。*
>
> *For the sake of our current vested interests, we can disregard the matter of **[Russians | Old Hairy]** for the time being.*

In Light ERE, only the fluent translations are tagged. In Rich ERE, however, both translation alternates are annotated and coreferenced when appropriate. In some cases, the fluent translations do not match the literal translations exactly in terms the entities, relations and events present. The annotation of the literal translations allows the English side to include annotations that may be present in the literal translation but not necessarily in the fluent translation. So, the annotation of the literal alternates should allow the English side to more exactly match the annotations on the Chinese side. This also supports Rich ERE's goal of more complete, exhaustive annotation.

5.2 Post Author Names

The data included some discussion forum metadata in the source text which contained, among other things the names of individual post authors, as in the example below.

```
<post author="服务咨询" datetime="2011-10-15T16:09:00" id="p5">
```

Post author names in the metadata were not translated/transliterated into English. These names typically consist of Chinese characters, Roman letters or a combination of the two.

In both Light and Rich ERE, the Chinese side annotated poster names and coreferenced them with mentions in the body of the message as appropriate. Because these poster names were frequently in Chinese characters and are uninterpretable to English annotators, they were not annotated on the English side.

6. Conclusions

The parallel Chinese-English ERE corpora described here are the first large-scale parallel corpora developed by LDC to have entities, relations and events annotation.

The data consists of approximately 100,000 words of Chinese discussion forum documents plus the equivalent amount of English that was the product of high quality manual translation. ERE annotation was performed independently on both the Chinese and the English sides. This independent annotation can be used to evaluate the accuracy of cross-language annotation projection methods.

The ERE annotation framework was developed to provide a resource for training data for various component tasks for knowledge base population in support of the DEFT program.

The creation of this data also included a new cross-lingual QC which allowed the correction of errors that would not have been captured through normal monolingual corpus-wide QC procedures. It also provided examples of some differences between the languages, both due to intrinsic linguistic differences between the languages (such as *pro-drop*) and the result of various translation artifacts, that resulted in imperfect matching on all layers of annotation.

The resources described in this paper have been distributed to performers in the DARPA DEFT Program and to participants in the NIST TAC KBP evaluations, and will be subsequently published in LDC's catalog, making them available to the broader research community.

7. Acknowledgements

This material is based on research sponsored by Air Force Research Laboratory and Defense Advanced Research Projects Agency under agreement number FA8750-13-2-0045. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory and Defense Advanced Research Projects Agency or the U.S. Government.

8. References

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. (2014). A Comparison of the Events and Relations Across ACE, ERE, TAC-KBP, and FrameNet Annotation Standards. ACL 2014: 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, June 22-27. 2nd Workshop on Events: Definition, Detection, Coreference, and Representation.

Ann Bies, Justin Mott, Seth Kulick, Jennifer Garland and Colin Warner. (2014). Incorporating Alternate

Translations into English Translation Treebank. In *Proceedings of the 9th Edition of the Language Resources and Evaluation Conference* (LREC 2014), Reykjavik, May 26-31.

DARPA. (2012). *Broad Agency Announcement: Deep Exploration and Filtering of Text (DEFT)*. Defense Advanced Research Projects Agency, DARPA-BAA-12-47.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, Ralph Weischedel. (2004). *Automatic Content Extraction (ACE) Program - Task Definitions and Performance Measures.* 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, May 24-30.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. (2011). *Building a Multilingual Named Entity-Annotated Corpus Using Annotation Projection*. Recent Advances in Natural Language Processing. 2011.

Jennifer Garland, Stephanie Strassel, Safa Ismael, Zhiyi Song, Haejoong Lee. (2012). Linguistic Resources for Genre-Independent Language Technologies: User-Generated Content in BOLT, In Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation, Istanbul, May 21-27.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. (2010). Overview of the TAC 2010 knowledge base population track. In Third Text Analysis Conference (TAC 2010).

Seth Kulick, Ann Bies, Justin Mott. (2014). *Inter-annotator Agreement for ERE Annotation.* ACL 2014: 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, June 22-27.

Linguistic Data Consortium. (2005). *ACE (Automatic Content Extraction) English Annotation Guidelines for Events Version 5.4.3*. ACE 2005 Events Guidelines

NIST. (2015). TAC Knowledge Base Population (KBP) 2015. http://www.nist.gov/tac/2015/KBP/

E. Riloff, C. Schafer and D. Yarowsky. (2002). Inducing information extraction systems for new languages via cross-language projection. In Proceedings of Coling 2002, Taipei, Taiwan.

Zhiyi Song and Stephanie Strassel. (2008). Entity Translation and Alignment in the ACE-07 ET Task. *LREC*, Marrakech.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant and Xiaoyi Ma. (2015). From Light to Rich ERE: Annotation of Entities, Relations, and Events. The 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015). *3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*.

Christopher Walker, Stephanie Strassel, Julie Medero and Kazuaki Maeda. (2006). *ACE 2005 Multilingual Training Corpus*. Linguistic Data Consortium, LDC Catalog No.: LDC2006T06.

Mengqiu Wang and Christopher D Manning. (2014). *Cross-lingual projected expectation regularization for weakly supervised learning.* Transactions of the Association for Computational Linguistics. 2014.

David Yarowsky, Grace Ngai, and Richard Wicentowski. (2001). *Inducing multilingual text analysis tools via robust projection across aligned corpora*. In Proceedings of the First International Conference on Human Language Technology Research, HLT '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Imed Zitouni and Radu Florian. (2008). *Mention Detection Crossing the Language Barrier*. Empirical Methods in Natural Language Processing. 2008.

Ayah Zirikly, Masato Hagiwara. (2015). *Cross-lingual Transfer of Named Entity Recognizers without Parallel Corpora*. Association for Computational Linguistics. 2015.