Entity Translation and Alignment in the ACE-07 ET Task

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

Abstract

Entities -- people, organizations, locations and the like -- have long been a central focus of natural language processing technology development, since entities convey essential content in human languages. For multilingual systems, accurate translation of named entities and their descriptors is critical. LDC produced Entity Translation pilot data to support the ACE ET 2007 Evaluation and the current paper delves more deeply into the entity alignment issue across languages, combining the automatic alignment techniques developed for ACE-07 with manual alignment. Altogether 84% of the Chinese-English entity mentions and 74% of the Arabic-English entity mentions are perfect aligned. The results of this investigation offer several important insights. Automatic alignment algorithms predicted that perfect alignment for the ET corpus was likely to be no greater than 55%; perfect alignment on the 15 pilot documents was predicted at 62.5%. Our results suggest the actual perfect alignment rate is substantially higher (82% average, 92% for NAM entities). The careful analysis of alignment errors also suggests strategies for human translation to support the ET task; for instance, translators might be given additional guidance about preferred treatments of name versus nominal translation. These results can also contribute to refined methods of evaluating ET systems.

1. Introduction

Entities -- people, organizations, locations and the like -- have long been a central focus of natural language processing technology development, since entities convey essential content in human languages. For multilingual systems, accurate translation of named entities and their descriptors is critical. For instance, [Babych 2003] reports that integrating a named entity recognition module into existing machine translation systems would improve MT system performance by 20%. The Automatic Content Extraction (ACE) Program develops extraction technology to support automatic processing of source language data, including classification, filtering, and selection based on the language content of the source data. In 2007 ACE extended its standard suite of evaluation tasks to include a pilot entity translation (ET) task. ET participants take Arabic or Chinese text as input, and output an English language catalog of all entities mentioned in those documents. The task is not limited to named entities, but also includes descriptors (nominal phrases) and pronouns across seven entity types: Persons, Organizations, Locations, GeoPolitical Entities, Facilities, Vehicles and Weapons. System performance is evaluated along a number of parameters, including coverage of the entities recognized as well as the quality of the English language renderings of each entity's mentions [NIST 2007]. Systems are not required to identify where in the Chinese or Arabic source document a given entity mention or temporal expression has been detected; this is in keeping with a desire to gradually move information extraction technology for ACE toward a knowledge base model where evaluation will consist of measuring the state of a database (knowledge base) after some amount of source text has been processed.

2. Production of the ACE-07 Entity Translation corpus

Linguistic Data Consortium creates linguistic resources -- annotated corpora, tools and best practices -- to support the ACE Program. For the ET pilot evaluation, LDC created annotated devtest and evaluation corpora in two genres: newswire and weblogs. Source data from each language is translated into the remaining two languages; so for instance, Arabic texts are translated into both Chinese and English. Devtest consists of 22.5Kw of source data per language, while eval consists of 15Kw per language. Prior to translation, the source data is manually segmented into sentence units (SU). Each segmented source document is translated by professional translators following guidelines developed by LDC for the DARPA GALE program [LDC 2006a, 2006b]. SUs are preserved during the translation, resulting in parallel text that is aligned at the segment level. Character offsets for each SU are retained during subsequent annotation and document processing.

After translation, both source and translated documents are annotated by LDC for entities and temporal expressions, following the standard ACE annotation task definitions [LDC2006c, Ferro2005]. Entity annotation includes labeling the extent, head, type, subtype, class (specific/generic) and level (name, nominal or pronoun) for every entity mentioned in the text, and co-referencing multiple mentions of the same entity within each document. Manual co-reference of entities across documents and languages is not performed. No attempt is made to manually align entity

mentions across documents/languages, although it is expected that the human translations result in comparable entity mentions within for the same segment across the three languages, given that translation is based on pre-defined SU segments.

	English Source	Chinese Translation	Arabic Translation
Segment ID	<en="7" start="536" end="614">	<cn="7" start="223" end="254">	<ar="7" start="483" end="591">
Source/ Translation	Carnahan, a Democrat, was 66. He had served as Missouri's governor since 1992.	来自民主党的卡纳汉现年66岁，1992年起担任密苏里州州长。	ويبلغ كارناهان القادم من الحزب الديموقراطي ستة وستين سنة من العمر وكان قد أصبح حاكما لولاية ميسوري سنة 1992.
English gloss	---	From Democratic Party DE Carnahan this year 66 years old, 1992 year since served as Missouri state governor.	The age of Carnahan coming from the Democratic Party is 66 years old, and he was the governor of the Missouri State 1992.

Table 1: An aligned translation segment with English gloss

To enable post-hoc analysis of human entity translation alignment (and to provide added insight into system performance), LDC created a Predicted Entity Alignment Table listing every entity mention for a given segment, across all three languages (source language plus two translations). The table is arranged by Document ID then by segment, and includes a unique EntityID plus information about entity type, subtype, head, level and class for every entity mention in each of the three languages. As noted, the entity mentions themselves are not manually aligned or mapped across languages. LDC did develop an alignment algorithm to enable further research [Walker 2007], but the model was error-prone and failed to reach optimal alignment, especially for non-named entities.

3. Entity Mention Alignment

The current paper delves more deeply into the entity mention alignment issue, combining the automatic alignment techniques developed for ACE-07 with manual alignment. The manual alignment task identified 15 Chinese-English and 15 Arabic-English translation documents pairs from the ET evaluation corpus. Starting with the Predicted Entity Alignment Table plus automatic alignment output, we sort entity mentions for each segment by type, subtype and mention level. This output is validated by human annotators to maximize possible matches. Each entity pair is then judged for its alignment status, using eight categories illustrated in the table below. Alignment pairs may be assigned to multiple categories.

Category	Explanation	Source/ Translation	Annotation
Perfectly aligned	A mention pair agrees in every aspect (head, type, subtype, mention level, mention class)	缅甸	(NAM, GPE-Nation, SPC)
		Myanmaran	(NAM, GPE-Nation, SPC)
Type changed	A mention pair differs in entity type	代表团	(NOM, PER, Group, SPC)
		delegation	(NOM, ORG, Non-Government, SPC)
Subtype changed	A mention pair differs in entity subtype	奉辛比克党	(NAM, ORG, Non-Government, SPC)
		Funcinpec Party	(NAM, ORG, Government, SPC)
Level changed	A mention pair differs in mention level	双方	(PRO, GPE, Nation, SPC)
		parties	(NOM, GPE, Nation, SPC)
Class changed	A mention pair differs in mention class	海军	(NOM, ORG, Government, SPC)
		navy	(NOM, ORG, Government, USP)
Mention split	A mention in source language is split into two mentions in the target language or vice versa	陆家嘴金融贸易区	(NAM, LOC, Region-general, SPC)
		Lujiazui	(NAM, GPE, County-or-district, SPC)
		District	(NAM, GPE, County-or-district, SPC)
Missing from source	An mention in translation lacks counterpart in source due to translation or annotation error	美元	美 is not taggable
		US (dollar)	(NAM, GPE, Nation, SPC)
Missing from target	An mention in source lacks counterpart in translation due to translation or annotation error	外国	(NOM, GPE, Nation, SPC)
		foreign	not taggable in English

Table 2: Manual entity alignment categories

3.1 Chinese-English alignment

Within the 15 document Chinese-English files, there are 786 Chinese entity mentions while the English translations of those documents contain a total of 820 entity mentions; 14 of the 786 Chinese mentions are missing from English, while 61 of the English mentions are missing from Chinese. Altogether there are 772 mentions pairs across Chinese and English and 84% of them are perfect aligned. Below is the distribution for each alignment category:

Category	mention pairs assigned to this category	% of total mention pairs
Perfectly aligned	646	84%
Type changed	31	4%
Subtype changed	25	4%
Level changed	14	2%
Class changed	41	5%
Mention split	6	1%
Total mention pairs	772	

Table 3: Manual alignment results for Chinese-English documents

We examine alignment within the three mention levels in more detail. Not surprisingly, of the 772 Chinese-English mention pairs, we find that named mentions (NAM) have the highest level of perfect alignment (92%). Nominal descriptors (NOM) are next with 64% perfect alignment. As for Pronominal entity mentions (PRO), there are altogether 16 pairs and 15 of them are perfectly aligned. However, the lot of the English PROs are missing their Chinese counterparts (23 out of 49). This is largely due to the added wh- connectors in English relative clauses and the necessary insertion of pronouns when translating into English from Chinese, which is a pro-drop language. Below is the distribution of each category in NAM, NOM and PRO:

Category	NAM	NOM	PRO
Perfectly aligned	439	192	15
Type changed	9	22	0
Subtype changed	15	9	1
Class changed	0	41	0
Level changed	6	8	0
Mention split	6	0	0

Table 4: Alignment categories across mention levels of Chinese-English documents

3.2 Arabic-English alignment

Within the 14 Arabic-English documents, there are 1182 Arabic entity mentions while the English translations of those documents contain a total of 1521 entity mentions; 101 of the 1182 Arabic mentions are missing from English while 430 of the 1521 English mentions are missing from Arabic. Altogether there are 1081 mentions pairs across Arabic and English and 74% of them are perfect aligned. Below is the distribution for each alignment category:

Category	mention pairs assigned to this category	% of total mention pairs
Perfectly aligned	796	74%
Type changed	69	6%
Subtype changed	27	2%
Level changed	97	9%
Class changed	105	10%
Mention split	16	1%
Total mention pairs	1081	

Table 5: Manual alignment results for Arabic-English documents

Arabic-English alignment within mention levels has similar distribution as Chinese-English. Of the 1081 Arabic-English mention pairs, named mentions (NAM) have the highest level of perfect alignment (87%). Nominal descriptors (NOM) are next with 64% perfect alignment. As for Pronominal entity mentions (PRO), there are altogether 68 pairs and 35 of them are perfectly aligned. Below is the distribution of each category in NAM, NOM and PRO:

Category	NAM	NOM	PRO
Perfectly aligned	410	351	35
Type changed	17	49	3
Subtype changed	12	14	1
Class changed	7	85	13
Level changed	23	59	15
Mention split	10	6	0

Table 6: Alignment categories across mention levels of Chinese-English documents

Compared to Chinese-English alignment, Arabic-English has many more missing mentions from either the target or the source language. For NAM and NOM, the missing mentions are primarily due to annotation inconsistency across the two languages. For PRO, Arabic has far less PRO mentions than English does, as possessive and object pronouns are attached to other words (for example house-my, leader-its, kill-him) while subject pronouns are most often dropped.

Category	NAM	NOM	PRO
mention pairs	471	552	68
missing from Arabic	21	110	299
missing from English	18	72	15

Table 7: Entity mention missing in Arabic-English documents

4. Conclusion

The results of this investigation offer several important insights. Automatic alignment algorithms predicted that perfect alignment for the ET corpus was likely to be no greater than 55%; perfect alignment on the 15 pilot documents was predicted at 62.5% [Walker 2007]. Our results suggest the actual perfect alignment rate is substantially higher (82% average, 92% for NAM entities). The careful analysis of alignment errors also suggests strategies for human translation to support the ET task; for instance, translators might be given additional guidance about preferred treatments of name versus nominal translation. These results can also contribute to refined methods of evaluating ET systems.

5. References

Babych B., Hartley A.. Improving Machine Translation quality with automatic Named Entity recognition. 2003. In: EACL 2003, 10th Conference of the European Chapter. *Proc. of the 7th Int. EAMT workshop on MT and other language technology tools*. Budapest Hungary (2003) pp. 1-8

Ferro, L., Gerber, L., Mani, I., Sundheim, B. and Wilson G. 2005. TIDES 2005 Standard for the Annotation of Temporal Expressions
http://timex2.mitre.org/annotation_guidelines/2005_timex2_standard_v1.1.pdf

NIST. 2007. The Evaluation Plan for the ACE 2007 Pilot Evaluation of Entity Translation. http://www.nist.gov/speech/tests/ace/ace07/doc/ET07-evalplan-v1.8.pdf

LDC. 2006a. GALE Arabic Translation Guidelines V2.3
http://projects.ldc.upenn.edu/gale/Translation/specs/GALE_Arabic_translation_guidelines_v2.3.pdf

LDC. 2006b. GALE Chinese Translation Guidelines V2.3
http://projects.ldc.upenn.edu/gale/Translation/specs/GALE_Chinese_translation_guidelines_v2.3.pdf

LDC 2006c. ACE English Entity Annotation Guidelines V5.6.6
http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v5.6.6.pdf

Day, D. Entity Translation 2007 Pilot Evaluation. Presentation to the ACE/ET Evaluation workshop, March 2007.

Walker, C. REFLEX ET Cross-Linguistic Agreement**.** Presentation to the ACE/ET Evaluation workshop, March 2007.