NTU Multilingual Corpus (NTUMC)

Overview:

- Ongoing development @ NTU, HSS;
- Open Source (CC BY); (but not yet openly available!)
- 4 languages: English, Mandarin, Japanese, Indonesian;
  - (+ Arabic, Korean, Vietnamese and Thai data)
- 4 main genres:
  - short-story (public domain), essay (open source), news and tourism;
- Aprox. 23,000 sentences (UTF-8 Encoding):

Table 1 - Size of the current (1st release) version of NTUMC

	Sentences				Word	Concepts
Genre	eng	cmn	jpn	ind	eng	eng
Story 1:	599	606	698	–	11,200	5,300
Story 2:	599	612	702	–	10,600	4,600
Essay:	769	750	773	–	18,700	8,800
News:	2,138	2,138	2,138	–	55,000	23,200
Tourism:	2,988	2,332	2,723	2,197	74,300	32,600
Total	7,093	6,438	7,034	2,197	169,800	74,600

NANYANG TECHNOLOGICAL UNIVERSITY

Release / Annotation

Release (by request): SQLITE Database (xml in the future)

Two Layers of annotation:

- Monolingual Layer:
  - Sentence Segmented;
  - Tokenised (relevant ‘continuous script’ languages);
  - POS tagged;
  - Sense tagged (using the Open Multilingual Wordnet);
- Cross-lingual Layer:
  - Sentence Alignment;
  - Sense Alignment;

Future Plans:

- Deep structural semantic and syntactic annotation;

NANYANG TECHNOLOGICAL UNIVERSITY

NTUMC - Usage

Cross-lingual distribution analysis of pronouns in ENG, CMN and JPN

- ENG > highest number of pronouns (2nd CMN, 3rd JPN);
- ENG has more translated counterparts in CMN (vs. JPN);
  - 26.9 % do not have a link in CMN;
  - 68.8% do not have a link in JPN;
- CMN has the highest proportion of 'contentful' pronouns;

Rates and patterns of Chengyu distribution across genres

- Idiomatic expressions often appear in Chinese where there was none in the source English text; (against predictions);
- Story genre has the highest % of types, but the lowest per sentence;
- Essay genre has the highest per sentence occurrence (~1 per 7);

NANYANG TECHNOLOGICAL UNIVERSITY

References

- Francis Bond, Shan Wang, Eshley Huini Gao, Hazel Shuwen Mok, and Jeanette Yiwen Tan. 2013. Developing parallel sense-tagged corpora with wordnets. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (LAW 2013). Sofia. pp 149–158.

- Wan Yu Ho, Christine Kng, Shan Wang and Francis Bond. 2014. Identifying Idioms in Chinese Translations In 9th Edition of the Language Resources and Evaluation Conference (LREC 2014), Reykjavik.

- Liling Tan and Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). In International Journal of Asian Language Processing 22(4) pp 161–174.

- Yu Jie Seah and Francis Bond. 2014. Annotation of Pronouns in a Multilingual Corpus of Mandarin Chinese, English and Japanese. In 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation Reykjavik.

- Shan Wang and Francis Bond. 2014. Building The Sense-Tagged Multilingual Parallel Corpus. In 9th Edition of the Language Resources and Evaluation Conference (LREC 2014), Reykjavik.

NANYANG TECHNOLOGICAL UNIVERSITY

ISLRN (Basic Metadata)

Title	NTUMC
Full Official Name	NTU Multilingual Corpus
Resource Type	Other, sense annotated multilingual parallel corpus
Source/URL	http://compling.hss.ntu.edu.sg/ntumc
Format/MIME Type	text/xml
Size/Duration	22,762 sentences
Format/MIME Type	text/plain
Access Medium	Download
Description	Sentence and sense-aligned parallel corpus of English, Mandarin Chinese, Japanese and Indonesian. It included POS and sense tagging at a monolingual layer. and cross-lingual annotation with sentence and sense alignments.
Version	1.0
Media Type	Text, Interactive Resource
Language(s)	English, Mandarin Chinese, Japanese, Indonesian
Resource Creator	Liling Tan and Francis Bond
Distributor	Francis Bond
Rights Holder	Liling Tan and Francis Bond