Evaluating Term Extraction Methods for Interpreters

Ran Xu, Serge Sharoff
Centre for Translation Studies
School of Modern Languages and Cultures
University of Leeds, UK,
{sml3rx,s.sharoff}leeds.ac.uk

Abstract

The study investigates term extraction methods using comparable corpora for interpreters. Simultaneous interpreting requires efficient use of highly specialised domain-specific terminology in the working languages of an interpreter with limited time to prepare for new topics. We evaluate several terminology extraction methods for Chinese and English using settings which replicate real-life scenarios, concerning the task difficulty, the range of terms and the amount of materials available, etc. We also investigate interpreters' perception on the usefulness of automatic termlists. The results show the accuracy of the terminology extraction pipelines is not perfect, as their precision ranges from 27% on short texts to 83% on longer corpora for English, 24% to 31% on Chinese. Nevertheless, the use of even small corpora for specialised topics greatly facilitates interpreters in their preparation.

1 Introduction

The study investigates term extraction methods using comparable corpora for interpreters. Simultaneous interpreting requires efficient use of highly specialised domain-specific terminology in the working languages of the interpreter. By necessity, interpreters often work in a wide range of domains and have limited time to prepare for new topics. To ensure the best possible simultaneous interpreting of specialised conferences where a great number of domain-specific terms are used, interpreters need preparation, usually under considerable time pressure. They need to familiarise themselves with concepts, technical terms, and proper names in the interpreters' working languages.

However, there is little research into the use of modern terminology extraction tools and pipelines for the task of simultaneous interpretation. At the start of computer-assisted termbank development, Moser-Mercer (1992) overviewed the needs and workflow of practicing interpreters with respect to terminology and offered some guidelines for developing term management tools specifically for the interpreters. That study did review the functionalities of some termbanks and term management systems, yet there was no mention of corpus collection (a fairly new idea at the time) or automatic term extraction.

A few previous studies mentioned the application of corpora as potential electronic tools for the interpreters. Fantinuoli (2006) and Gorjanc (2009) discussed the functions of specific online crawling tools and explored ways to extract specialised terminology from disposable web corpora for interpreters. Our work is most closely connected to Fantinuoli's work on evaluation of termlists obtained from Web-derived corpora. However, that study relied on a single method of corpus collection and term extraction, and did not include an investigation into integration of corpus research into practice of interpreter training.

Rütten (2003) suggested a conceptual software model for interpreters' terminology management, in which termlists are expected to be extracted (semi-)automatically and then to be revised by their users, the interpreters, who can concentrate on those terms which are relevant and important to remember. However the study neither tested the functions of the term extraction tools nor further discussed interpreters' perception on the usefulness of the automatically lists in their preparation for interpreting tasks.

Proceedings of the 4th International Workshop on Computational Terminology, pages 86–93,
Dublin, Ireland, August 23 2014.

	FR0		FR1		FR2		SM1		SM2	
	En	Zh	En	Zh	En	Zh	En	Zh	En	Zh
Texts	1	1	9	9	81	86	9	12	74	84
Size	774	1,641	42,006	30,174	206,197	129,350	20,533	40,545	166,499	116,235

Table 1: Corpora used in this study (the size is in words for En, in characters for Zh)

Based on Rütten model, this paper will further test the functions of several term extraction tools for English and Chinese, and will discuss the interpreters' perception on the usefulness of the automatically-generated lists in their preparation for interpreting tasks.

In the remainder of this paper we will describe the pipelines for corpus collection and terminology extraction (Section 2), present the results of their numeric evaluation (Section 3), and discuss options for future research, including the challenges for the term extraction pipelines in this setting (Section 4).

2 Corpus collection and term extraction

In this section we describe pipelines for interpreters' terminology preparation with the use of term extraction tools. We compare several approaches to corpus compilation and processing for specialised texts as well as several pipelines for terminology extraction.

2.1 Description of the procedure

Two specialized topics In this study MA student interpreters were invited to prepare for simultaneous interpreting tasks on two specialised topics: fast reactors (FR) and Seabed minerals (SM). They were provided with two monolingual specialised corpora in both English and Chinese for their advance preparation on each of the topics (FR & SM).

Three term extractors The students started with the FR topic, and were asked to manually generate their own lists from the provided corpora (FR1 En & Zh) before their simultaneous interpreting exercise on the topic in both directions (English→Chinese and Chinese→English). After their interpreting tasks, they were then asked to evaluate the relevance of two monolingual lists (En & Zh) which were automatically generated by one of the three tools (TTC TermSuite, Syllabs Tools and TeaBoat). The purpose here is to see which tool could extract more relevant terms for the needs of trainee interpreters.

We collected and compared the annotation results from the students to select a single tool with comparatively better performance. We then invited the students to prepare for the other topic (SM) by using automatically-generated lists in the simultaneous interpreting preparation.

2.2 Corpus compilation

There are two types of sources where comparable corpora are from:

1. Conference documents and relevant background documents provided by the conference organisers
2. Specialised corpora collected from the internet using WebBootCat (Baroni and Bernardini, 2004)

Table 1 presents all the corpora we use in this study. FR0/SM0 has been created from a single relevant document, representing the speech that the trainee interpreters were asked to interpret from in this experiment. We also ran term extraction from this "corpus" since often a text of this length is the only source of information given to the interpreters in advance. We tried to balance the terminological difficulty for both languages, even if this was not always possible. After manual term selection, we found that FR0-Zh contains 147 terms per 591 seconds of delivery (15 terms per minute), FR0-En: 86 terms per 566 seconds (9 t/min), SM0-Zh: 157 terms per 604 seconds (16 t/min), SM0-En: 169 terms per 750 seconds (14/min).[1]

[1] Counting the term density per unit of text is not straightforward, because of very different tokenisation rules in Chinese and English.

Seeds (En)	**Seeds (Zh)**
fast breeder reactor	快中子增殖反应堆
fission	裂变
decay heat	余热
uranium	铀
plutonium	钚
core damage	堆芯损坏
Fukushima accident	福岛事故
nuclear waste	核废料
fuel cycle	燃料循环
coolant	冷却剂

Table 2: Parallel keyword seeds on Fast Reactors for FR2

FR1 (En & Zh) and SM1 (En & Zh) are comparable corpora, which represent conference documents and relevant background documents passed from the conference organisers, including speech outlines, research papers from experts and research institutes, reports from national and international authorities, as well as popular science articles, Wikipedia articles, specialised journal articles and interviews, etc.

FR2 (En & Zh) and SM2 (En & Zh) are corpora collected by Web crawling using Bootcat(Baroni and Bernardini, 2004). For instance, to produce FR2 we started with a set of ten relevant keywords in English and Chinese as shown in Table 2, then used BootCat to retrieve online resources and generate two corpora (FR2 En & Zh). All the keyword seeds are from the English speech-FR0 that the students were going to interpret from, and are therefore considered very relevant and important terms. The Chinese keywords are the translations of the English ones.

Preprocessing included webpage cleaning (Baroni et al., 2008), as well as basic linguistic processing. Lemmatisation and tagging for English was done using TreeTagger (Schmid, 1994), while for Chinese we used "Segmenter", an automatic tokenisation tool (Liang et al., 2010) followed by TreeTagger for POS tagging. Lemmatisation is needed because the keywords in a glossary are expected to be in their dictionary form. Lemmatisation also helps in reducing the nearly identical forms, e.g., *sulphide deposit*(*s*). However, lemmatisation also leads to imperfect terms, e.g., *recognise type of marine resource*, while the plurals and participles should be expected in a dictionary form (*recognised type of marine resources*).

2.3 Automatic term extraction

TTC TermSuite (Daille, 2012) is based on lexical patterns defined in terms of Part-of-Speech (POS) tags with frequency comparison against a reference corpus using specificity index (Ahmad et al., 1994), which extracts both single (SWT) and multi-word terms (MWT) outputs their lemmas, part of speech, lexical pattern, term variants (if any), etc. The most important feature of the TTC TermSuite is the fact that term candidates can be output with their corresponding term variants. Syllabs Tools (Blancafort et al., 2013) is a knowledge-poor tool, which is based on unsupervised detection of POS tags, following the procedure of (Clark, 2003), and on the Conditional Random Field framework for term extraction (Lafferty et al., 2001). Teaboat (Sharoff, 2012) does term extraction by detecting noun phrases using simple POS patterns in IMS Corpus Workbench (Christ, 1994) and by applying log-likelihood statistics (Rayson and Garside, 2000) to rank terms by their relevance to the corpus in question against the Internet reference corpora for English and Chinese (Sharoff, 2006).

3 Term extraction evaluation

Fantinuoli (2006) used five categories to find the level of specialisation and well-formedness of the automatically-generated candidate termlist:

1. specialised terms that were manually extracted by the terminologist (and are contained in the reference term list);

FR-TTC		FR-Teaboat		FR-Syllabs		SM-Syllabs	
EN	ZH	EN	ZH	EN	ZH	EN	ZH
0.541	0.500	0.166	0.435	0.181	0.662	0.117	0.221

Table 3: Krippendorff's α for different term lists

2. highly specialised terms that were not detected by the terminologist;
3. non-specialised terms that are commonly used in the field of his study (medicine);
4. general terms that are not specific to the medical field;
5. ill-formed, incomplete expressions and fragments.

Our annotation system extends Fantinuoli's study because the purpose of annotation in this project is to give the interpreters possibility to extract relevant terms from all the candidate terms regardless of their levels of specialisation. Our premise is that interpreters may need relevant terms, both highly specialised and less specialised, in order to prepare themselves for a conference. The annotators are the end users of the list, i.e. the trainee interpreters who participated in this research. Since the interpreters are tasked with translating speeches in the domain, they need themselves to decide what is likely to be relevant instead of relying on the terminologists who describe the overall structure of the domain. The following is the five-category annotation system that we used in this research:

R relevant terms (terms closely relevant to the topic), eg. *breed ratio, uranium-238, decay heat removal system*;

P potentially relevant terms (a category between "I" and "R": they are terms; but annotators are not sure whether they are closely relevant to the topic of their assignment), eg. *daughter nuclide, neutron poison, Western reactor*;

I irrelevant terms (terms not relevant to the topic), eg. *schematic diagram, milk crate*;

G general words (rather than terms), eg. *technical option, monthly donation, Google tag, discussion forum*;

IL ill-formed constructions (parts of terms or chunks of words), eg. *var, loss of cooling, separate sample container, first baseline data, control ranging*.

It only took several minutes to generate a termlist after uploading the designated corpus onto TTC TermSuite, Syllabs Tools and TeaBoat. Each of them automatically generated corresponding monolingual termlists sorted by their term specificity scores. For all the tools we set the threshold of obtaining 500 terms (if possible), as a practical limit for all evaluation experiments.

The trainee interpreters were asked to annotate the list by using the above annotation system. Each of them reported that it took them about 60 minutes to annotate both lists (in EN & ZH) on each of the topics (FR & SM). All the annotators were briefed about what counts as terms and the annotation system before they started their evaluation of term lists. We aim for consistency, yet inter-annotator disagreement does exist and there is a certain degree of subjectivity in annotation. To measure the level of agreement we used Krippendorff's α over the other measures, such as Fleiss' κ, because Krippendorff's α offers an extension of such measures as Fleiss' κ and Scott's π by introducing a distance metric for the pairwise disagreements, thus making it possible to work with interval-scale ratings, e.g., considering disagreement between **R** and **P** as less severe than between **R** and **I** (Krippendorff, 2004).

The values of Krippendorff's α (see Table 3) are relatively low. The most common cases of disagreement are between **R** and **P** (the boundary between them often depends on the amount of knowledge on the side of the annotator), but also quite surprisingly between **R** and **IL**, when some annotators interpret ill-formed sequences as a contribution to useful terms.

With the disagreement taken into consideration, our evaluation on the number of relevant terms was judged by the agreement between at least two annotators among four to six annotators for the topic of FR. This established the gold standard lists reported in Table 4.

The annotation results from Table 4 for English show that Syllabs generated more relevant terms than the other two tools from both FR0 and FR1. Both Syllabs and Teaboat generated good numbers of

	Tool	FR0	FR1	FR2	SM1
English:	Syllabs	85/104(82%)	309/500 (62%)	400/500 (80%)	441/500 (88%)
	Teaboat	44/56(79%)	232/376 (62%)	413/499 (83%)	
	TTC	NA	136/500 (27%)	287/500 (57%)	

	Tool	FR1	SM1
Chinese:	Syllabs	156/500(31%)	130/500 (26%)
	Teaboat	141/450(31%)	
	TTC	119/500(24%)	

Table 4: Number of relevant (R) terms against candidate terms

relevant terms from FR2. In addition, Syllabs' and TeaBoat's English lists contain more specialised terms in the domain of FR, such as *defence-in-depth, once-through fuel cycle, suppression chamber of the containment*, etc. These specialised terms with relatively low frequency are not included in the TTC's list. The terms included in TTC's list are more general terms, such as *steam, energy, liquid, heat, leak*, etc., which are likely to be already known by the trainee interpreters.

The English termlists from all the tools contain a number of repetitions in the form of term variants, following Daille's definition as "an utterance which is semantically and conceptually related to an original term" (Daille, 2005). The automatically generated termlists contain the following types of term variations, which are counted as individual term candidates scattered in the termlists:

Morphological variation: *bathymetry* vs *bathymetric* (not different when translated into Chinese)

Anaphoric variation: *pollymetallic sulphide deposit* vs *deposit*

Pattern switching: *meltdown of the core* vs *core meltdown*; *level of gamma radiation* vs *gamma radiation level*

Synonymy in variation: *deep sea mining* vs *deep seabed mining*, *seabed* vs *seafloor*, *ferromanganese crust* vs *iron-manganese crust*

One the one hand, these variations provide useful lexical information about the term, preparing the interpreters for what is possible in their assignment; on the other hand, the term variants need to be explicitly linked, which is possible only in the TTC TermSuite tool.

The annotation results from Table 4 for Chinese show, both Syllabs' and Teaboat's lists offer obviously less relevant terms from FR1 compared with the English lists. When we further investigate the distribution of the term classes in annotations in Table 5, Syllabs' Chinese list on FR1 contains a large number of ill-formed constructions, including incomplete terms, eg. 水堆 'water reactor', 里岛核电站 'Mile Island nuclear plant' and longer chunks, eg. 最大程度上保证了钠, 可用压水堆后处理得到的钚作为核燃料. Teaboat's list contains a number of general words, eg. 开发 'development', 生产 'production' or 工程 'project'. Both categories (G and IL) are frequent in the TTC's Chinese list.

On the basis of these results, we selected a single tool (Syllabs) with comparatively better performance in both languages to generate termlists on SM1 (En & Zh) and asked 12 annotators to select the relevant terms and learn the terms during their interpreting preparation. Among the 500 candidate terms for English, 441 terms were agreed as relevant by at least two annotators, 266 terms were agreed by five annotators. Precision rates are 88.2% and 53.2% respectively. On the other hand, only 130 terms were agreed as relevant by two annotators from the 500 Chinese candidate terms. The precision rate for the Chinese list is 26%. The results basically replicate the previous findings on FR1.

The other pattern we observe from the current data is that the larger the corpus is, the more relevant terms the tools can generate. If the corpus is of very limited size (eg. FR0-en has only 774 words), the TTC TermSuite fails to generate any list for a 'corpus' of only 774 words, while the Syllabs and Teaboat tools produce shorter lists of 104 or 56 terms respectively. The situation is similar to other studies which used small (single-document) corpora, e.g., (Matsuo and Ishizuka, 2004).

	FR1-en	**FR2-en**	**FR1-zh**
Syllabs	500	500	500
R	309	400	156
P	90	53	73
I	15	10	5
G	56	16	46
IL	30	21	220
Teaboat	376	499	450
R	232	413	141
P	33	20	61
I	19	5	7
G	73	29	191
IL	19	32	50
TTC	500	500	500
R	136	287	119
P	48	1	32
I	3	1	4
G	310	205	209
IL	3	6	136

Table 5: Distribution of term annotation classes

4 Conclusions and future work

Reliability of the three term extractors The results show the accuracy of the terminology extraction pipelines is not perfect, as its precision ranges from 27% on short texts to 83% on longer corpora for English, 24% to 31% for Chinese. Among the three term extractors (TTC TermSuite, Syllabs Tools and Teaboat), Syllabs is more reliable in generating more relevant terms in English. All the three tools perform less satisfactory in generating relevant terms in Chinese. We hypothesise that at least three factors play an important role here:

1. Chinese is written without explicit word boundaries, while term extraction starts with already tokenised texts. Errors of the tokenisation process lead to difficulties in obtaining proper terms, e.g., 一回路 ‘primary loop’ becomes 一回 ‘once’ 路 ‘road’, also 和非能动安全性 ‘and passive security’ becomes 和非 ‘and not’ 能动 ‘active’ 安全性 ‘security’, which reduces the chances of detecting 非能动安全性 ‘passive security’ as a term.
2. Word ambiguity in Chinese is high. This leads to POS tagging errors, for example, when nouns are treated as verbs, and this breaks the POS patterns for term extraction, e.g., 示范堆 ‘demonstration reactor’ is treated as 示范/vn 堆/v.
3. Chinese exhibits more patterns than captured by the three term extraction tools we tested. For example, 并网发电 ‘connect to the grid’ is potentially a useful term, which is correctly POS-tagged as 并网/v 发电/vn, but not captured by the patterns in all the tools.

Two of the three causes of the results in Chinese concern text pre-processing. . Further investigation might be helpful in finding out how the pre-processing steps affect the performance of the term extractors and which terms are affected by each source of errors.

Manual selection Vs Automatic extraction of terms For the interpreters, manually selecting terms from a single document of limited size (eg. FR0-en=774 words) is possible. However, when conference documents amount to the size of FR1 (FR1-en=42,006 words), it took the trainee interpreters 9 hours on average to extract terms manually and to produce initial termlists, since they had to spend the majority of their time on reading through fairly complex documents, copying the terms from the texts onto their own termlists and searching for unfamiliar terms.

With the use of automatically-generated termlists on the same preparation task, students in the experiment group spent an average of 4 hours producing their initial bilingual termlists. Therefore half of the time spent on reading could be saved for the interpreters to get familiar with the concepts relevant to the terms and further activate the terms for their simultaneous interpreting tasks.

Furthermore, if interpreters are given limited time for preparation, they would not be able to read through larger corpora of the size of FR2 (FR2-en=206,197 words) and to produce termlists from them manually. That is probably when such tools we discussed in this article may have obvious advantage over the manual terms extraction by the interpreters. Moreover, in other studies we also demonstrated that in addition to providing an automatically-extracted termlist, it is also beneficial to link the terms to their uses in the concordance lines of the corpus they have been extracted from. This is expected to give the interpreters an easy access to the context of the terms to see how they are used and get more background knowledge about the domain.

Feedback from students After doing annotation, the students offered their written feedback on the termlists generated by the three term extractors. They also commented on the usefulness of the Syllabs' lists for their interpreting preparation.

They generally reported that the termlists provided many relevant terms on the two topics, and the use of the lists saved their precious preparation time. Some of them found the lists 'unexpectedly accurate and complete', and the presence of irrelevant words in the lists and the repetitions in the lists 'tolerable' (even taking into account the 24% to 31% precision rate for Chinese).

The students told us they used the lists as an important indicator for the content of the conference documents and relevant background documents. The lists helped them prioritise their preparation on the most relevant terms and concepts. Most of them expressed the opinion that if they are given very limited time, they would prefer to use the automatically-generated lists for their preparation. On the other hand, students reported that the termlists in Chinese offered much less relevant terms and contained quite a number of ill-formed constructions compared with the lists in English; therefore they felt the lists in Chinese were less useful and less reliable.

Extraction of proper names Proper names (including names of organisations, names of places, names and titles of people) are equally if not more important than terms for interpreters, yet many of them are not included in the automatically-generated lists by the three term extractors (TTC, Syllabs and Teaboat). Therefore, named entity extraction tools in addition to term extraction are needed to generate more complete lists for interpreters' use. This would be further explored in our future research.

File formats, plain text, encodings All the tools we tested can only process plain text (including UTF-8). Nevertheless, all the meeting documents are normally in one of the word processing formats (.pdf, .doc, .xls or .ppt) other than .txt. Interpreters need to take some time to convert all the files they obtain from their customers into plain text before they can possibly use any tool mentioned above.

References

Ahmad, K., Davies, A., Fulford, H., and Rogers, M. (1994). What is a term? The semi-automatic extraction of terms from text. In Hornby, M. S., Pöchhacker, F., and Kaindl, K., editors, *Translation studies: an interdiscipline*, pages 267–278. Amsterdam: John Benjamins Publishing Company.

Baroni, M. and Bernardini, S. (2004). Bootcat: Bootstrapping corpora and terms from the web. In *Proc. of LREC2004*, Lisbon.

Baroni, M., Chantree, F., Kilgarriff, A., and Sharoff, S. (2008). Cleaneval: a competition for cleaning web pages. In *Proc. of the Sixth Language Resources and Evaluation Conference, LREC 2008*, Marrakech.

Blancafort, H., Bouvier, F., Daille, B., Heid, U., Ramm, A., et al. (2013). TTC Web platform: from corpus compilation to bilingual terminologies for MT and CAT tools. In *Proceedings, Conference'Futures in technologies for translation (TRALOGY II)'*.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *COMPLEX'94*, Budapest.

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In *Proceedings of EACL*, pages 59–66.

Daille, B. (2005). Variations and application-oriented terminology engineering. *Terminology*, 11(1):181–197.

Daille, B. (2012). Building bilingual terminologies from comparable corpora: The TTC TermSuite. In *5th Workshop on Building and Using Comparable Corpora at LREC 2012*.

Fantinuoli, C. (2006). Specialized corpora from the web and term extraction for simultaneous interpreters. In Baroni, M. and Bernardini, S., editors, *WaCky! Working papers on the Web as Corpus*, pages 173–190. Gedit, Bologna. http://wackybook.sslmit.unibo.it.

Gorjanc, V. (2009). Terminology resources and terminological data management for medical interpreters. In Andres, D. and Pöllabauer, S., editors, *Spürst Du, wie der Bauch rauf-runter? Fachdolmetschen im Gesundheitsbereich*. `http://www.uni-graz.at/06gorjanc.pdf`.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3).

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML-01*.

Liang, M., Li, W., and Xu, J. (2010). *Using corpora: a practical coursebook*. Foreign Langue Teaching and Research Press, Beijing.

Matsuo, Y. and Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169.

Moser-Mercer, B. (1992). Banking on terminology conference interpreters in the electronic age. *Meta: Translators' Journal*, 37(3):507–522.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *Proc. of the Comparing Corpora Workshop at ACL 2000*, pages 1–6, Hong Kong.

Rütten, A. (2003). Computer-based information management for conference interpreters-or how will i make my computer act like an infallible information butler? In *Proc. Translating and the computer*, pages 14–14.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester.

Sharoff, S. (2006). Open-source corpora: using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.

Sharoff, S. (2012). Beyond Translation Memories: Finding similar documents in comparable corpora. In *Proc. Translating and the Computer Conference*, London.