DOI: 10.18721/JHSS.12209
УДК 8'33

SEARCHING FOR MULTICOMPONENT TERMS IN COMPARABLE SCIENTIFIC CORPORA

L.N. Beliaeva, O.N. Kamshilova

Herzen State Pedagogical University of Russia,
St. Peterburg, Russian Federation

The paper suggests the use of full-text parallel/comparable corpora with a "built-in" part of machine translation (MT) results for term extraction, harmonization and translation, since analysis and comparison of these texts will assure the possibility to identify terminological units for dictionary entries. We focus on the complicated and non-parallel structure of English multicomponent terminological noun phrases (NPs), their variants and modifications within the same text, which determine the need for a three-part text corpus, including parallel/comparable texts and their MT translation. The research has proved that multicomponent terminological NPs are not only specific for a scientific text, but they demonstrate ambiguous dependency relations, caused by their syntactic compression, which normally is the result of a sentence or of another NP convolution. These modifications are results of a number of standard procedures described in the paper.

Keywords: comparable corpora, MT, multicomponent NPs, terminological NPs, lexicography, noun phrase transformation.

Citation: L.N. Beliaeva, O.N. Kamshilova, Searching for multicomponent terms in comparable scientific corpora, Society. Communication. Education, 12 (2) (2021) 118–124. DOI: 10.18721/JHSS.12209

ПОИСК МНОГОКОМПОНЕНТНЫХ ТЕРМИНОВ В СОПОСТАВИМЫХ КОРПУСАХ НАУЧНЫХ ТЕКСТОВ

Л.Н. Беляева, О.Н. Камшилова

Российский государственный педагогический университет им. А.И. Герцена,
Санкт-Петербург, Российская Федерация

В статье предлагается использование полнотекстовых сопоставимых корпусов научных текстов со встроенной частью в виде выровненных результатов машинного перевода (МП). Такой корпус предназначен для решения задач извлечения, гармонизации и перевода терминологии, поскольку анализ и сравнение этих текстов позволяет идентифицировать терминологические единицы для формирования словарных статей. Особое внимание уделяется сложным и непараллельным структурам английских многокомпонентных терминологических именных групп, их вариантов и модификаций в рамках одного текста, что определяет необходимость трехчастного корпуса текстов, включающего параллельные/сопоставимые тексты и их машинный перевод. Исследование подтвердило, что многокомпонентные терминологические именные группы не только характерны для научных текстов, но демонстрируют многозначные отношения зависимостей, вызванные их синтаксической компрессией, что как правило является результатом свертки предложения или именной группы. Эти модификации в свою очередь являются результатом стандартных процедур, описанных в статье.

Ключевые слова: сопоставимые корпуса текстов, МП, терминологические именные группы, лексикография, трансформации именных групп.

Ссылка при цитировании: Beliaeva L.N., Kamshilova O.N. Searching for multicomponent terms in comparable scientific corpora // Society. Communication. Education. 2021. Vol. 12. No. 2. Pp. 118–124. DOI: 10.18721/JHSS.12209

Introduction

It has become almost commonplace, that much of terminographic job today is based on text corpora, that provide a reliable database for dealing not only with research issues, but with practical lexicographic tasks as well, such as terms identification and extraction, translation, etc., since corpus methods provide reliability and validity of empirical data [Beliaeva 2009, 2014; Beliaeva, Chernyavskaya, 2016; Delpech, Daille 2010; Heja 2010; TTC Project[1]]. Written text corpora, as a rule, include texts as they are, as well as text layouts: format boundaries and features, parsing results that are necessary for defining morphological characteristics of lexical units. Parallel and comparable text corpora are effectively used for creating multilingual lexicons and concordances.

We dare to suggest that, if we use full-text comparable corpora as a lexicographic base, it is necessary to expand them with a corpus of machine translation (MT) results. Analysis and comparison of findings in the comparable parts of corpora with the MT of the source part will make it possible to identify lexical units as candidates for special dictionary entries [*cf.*: Delgado et al. 2002; David, Curran 2007; Lavie et al. 2008]. The main difficulty in this identification process is to establish the boundaries and structures of these lexical units in a sentence and a text as a whole. Because scientific texts have an abundance of noun phrases (NPs) which are usually multiword units with a number of attributive elements modifying the head noun.

NPs have often been objects of study in both theoretical and applied aspects [Baroni, Zamparelli 2010; Bergsma, Wang 2007]. Though such phrases are functionally equivalent to a word, they actually represent a convolution of a sentence, i.e., they are definitely units of syntax, not lexicon, and their implicit dependency structure has always been a major issue for MT or human translation of scientific texts (cf.: [Feldman, Dagan 1995; Babych, B., Hartley 2002; Shen et al. 2008; Reiter, Frank 2010]), especially when translating from English to any inflectional language. The paper focuses on English multicomponent terminological NPs and candidates for their Russian equivalents.

Problem Statement

Applied lexicography (terminography) traditionally aims at building and updating subject oriented databases and special (terminological) automated/automatic dictionaries. The level and reliability of information that can be extracted from texts of various composition, structure and destination is determined by the lexicographic systems used for the purpose, their completeness and adequacy. The aim of the paper is to propose a way to optimize the use of comparable corpora as lexicographic resources by means of including a special part of MT results.

Methods

The research is based on corpus methodology, namely the principles of corpus building and balancing. The observations discussed in the paper have been made on corpus findings in original special research text corpora of different subject areas. The illustrations in the paper are from the two original corpora:

a) “Seismic Protection” corpus, a 1-million-word partly parallel corpus, the size of English and Russian parts is 500 000 tokens each;

[1] http://www.ttc-project.eu/about-ttc/concept-and-objectives

b) "Web and Linguistic Technologies" corpus, a comparable corpus, the size of English and Russian parts is approximately 230 000 tokens each.

NP interpretation and establishing a procedure for NP structure analysis also involve such methods as syntactic modelling, MT interpretation and comparison.

Results and Discusssion

An NP in its simplest form consists of one noun and its determiner. A muilticomponent NP includes a number of embedded premodifiers that make it a complex unit. NPs with a number of premodifiers are called simple if they include no preposition, no matter how many premodifiers they have [Malakhovskaya et al. 2021]. A multicomponent terminological NP actually represents a sentence compression (convolution). Its internal structure, consequently, must reflect the corresponding sentence structure, thereby revealing the syntactic dependencies. To recognize this structure in a concise form of an NP becomes a key problem, since the markers of relations between its actual components, which normally show in inflectional languages, hardly show in a simple English NP. The compression of sentence structure, the external simplification of both structure and form of English NPs cause semantic complication.

Our research in scientific text corpora of different subject areas (medicine, space systems, seismic isolation, power plants construction, machine translation, language teaching, etc.) demonstrate that 2-component NPs are the most frequent in English three times exceeding 3-component combinations, which are second frequent combinations in such texts (see Table 1).

Table 1. Distribution of English NP Models in "Web and Linguistic Technologies" corpus

Model number	Model	Length	Frequency	Number of different NPs
1	A+N	2	1474	748
2	A+PII+N	3	9	6
3	N1+N2	2	1407	530
4	N1+N2+N3	3	248	128
5	A/N1+N2	2	71	47
6	N1+G/N2	2	24	18
7	A1+A2/N1+N2	3	10	10
8	A+G/N2 +N2	3	7	4
9	A1+A2+N	3	151	104
10	A+N1+N2	3	292	172
11	PII+A+N	3	25	20
12	PII+N	2	170	73
13	A1+N1+A2/N2+N3	4	3	2
14	A1+C+A2+N	4	15	9
15	A1+A2/N1+N2+N3	4	6	7

The external simplicity of the most frequent English NPs is misleading as it is mostly the result of another NP or a sentence compression, which, as above mentioned, leads to its semantic complication.

According to this, we find two principal ways of constructing new NPs in a text:

a) either by adding a word to a standard or previously used NP, thus producing a novel, more complicated nomination:

machine translation => machine translation method, *machine translation service*, *machine translation program*, or

b) by deleting implicitly obvious units, thus condensing the sentence structure to a multicomponent NP: *syntactic dependency*, *syntactic formalism*, *syntactic dependency tree annotation => dependency annotation formalism*,

where (a) is a step-by-step process of gradual conversion – complication, adding specific characteristics to the head element, while (b) presents a conversion process of sequential convolution.

The cases of NP standard modifications considered above do not show all possible variants of NP development in a text. However, they might be helpful for extraction, harmonization and translation of NPs with a high degree of structure compression in parallel or comparable corpora of a particular subject area. Searching for a Russian equivalent of an English multicomponent term in a comparable corpus may be effectively supported by an MT stage.

Dealing with NP complicated structure, its identification and translation, we find only two approaches which can be used both in MT and human translation. The first approach includes modelling the knowledge base of the domain in question (within the framework of the MT system) or appealing to the translator's experience and their subject knowledge. In the case of MT this approach involves extensive research into possible relationships between the domain basic concepts and the items of the linguistic database. That actually means creating a semantic net, which is not only extremely laborious, but also space-consuming. Moreover, sometimes it is impossible to achieve an unambiguous solution to the problem. For example, a semantic network for *constant amplitude deformation cycle* would show relations between the nodes *constant* and *amplitude*, *constant* and *deformation*, *constant* and *cycle*, and this information is helpless both for MT and human translation as it doesn't explain the NP dependency structure.

The second approach is more formal and appropriate: we obtain the information from the entire text analysis. The procedure is based on formal indicators of the author's intentions, which are reflected both in the text structure and in the composition of different NPs with the same components.

Concordancing in scientific text corpora provides NP contextual analysis within the text space and leads to establishing procedures of coining novel NPs from those featuring in the text and to recognizing the compressed sentence structure in a concise form of an NP.

To establish the procedures, we suggest to use MT results for the source part of a parallel corpus as a reference base. Thus, comparing the machine translations of original English texts and the comparable Russian texts in a corpus, we can find exact matches of NPs, as well as partly matching NPs and terms presented in full and compressed structures.

Thus, for instance, the term *fatigue* in several NPs is presented in the text fragment *The detail category is the numerical designation given to a particular detail for a given direction of stress fluctuation, in order to indicate which* ***fatigue strength curve*** *is applicable for the* ***fatigue assessment*** (*The detail category number indicates the* ***reference fatigue strength*** $\Delta\sigma_C$ *in N/mm²*).

Analysis of texts across different subject areas has shown, that if an NP of more than two components appears in the text, it is generally followed by a 2-component NP in the nearest context within the limits of 2–3 sentences, or it can be found in the title, keyword list or abstract. Hence, in human translation, this fact can be a clue for NP structure diagnostics. Searching for parallel corpora, we may fail to fix such relations, but by referring to MT results as a storage base, we can optimize term identification and translation.

For instance, a 3-component NP *design equipment models* in the source English part can be variably translated as *модели расчетного оборудования* or *расчетные модели оборудования*. The English part has also a 2-component NP *design models*, its MT is *расчетные модели*, which finds an exact match in the Russian part: *расчетная модель*. But there is no variant of *design equipment* with an expected MT *расчетное оборудование*. Nothing similar is found in the Russian part, either. The comparison suggests that *design models*/*расчетные модели* demonstrates stronger dependences between *design* and *models* in

the texts of this subject area, than between *design* and *equipment*. So, the right candidate for a dictionary entry is *расчетные модели оборудования*.

Thus, corpus-based terminographic work may be improved by applying MT procedures and results to fix and store the history of NP conversion and modifications in the corpus.

Conclusion

The research has proved that multicomponent terminological NPs are not only specific for a scientific text, but they demonstrate ambiguous dependency relations, caused by their syntactic compression, which normally is the result of a sentence or of another NP convolution.

We argue that searching for a terminological NP variants and modifications within the same text or texts of the subject area helps to establish its dependency relations. These modifications are results of a number of standard procedures described in the paper.

Exploring parallel and comparable text corpora for terminological equivalents usually finds few exact matches and the NP modifications may show no evident likeness of their components. Corpus technics for term extraction, recognition and harmonization in two-part parallel and comparable corpora can be developed by the proposed decision, namely by adding an MT results corpus as a reference base.

REFERENCES

[1] **B. Babych, A. Hartley,** Improving machine translation quality with automatic named entity recognition. In Proceedings of the 7-th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT. Association for Computational Linguistics, 2003, pp. 1–8.

[2] **M. Baroni, R. Zamparelli,** Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 1183–1193.

[3] **L. Beliaeva,** Applied Lexicography and Scientific Text Corpora, Transcations on Business and Engineering Intelligent Applications. Galina Setlak, Kassimir Markov (ed.). Rzeshov, Poland: ITHEA, 2014, pp. 55–63

[4] **L. Belyaeva,** Scientific Text Corpora as a Lexicographic Source, SLOVKO 2009. NLP, Corpus Linguistics, Corpus Based Grammar Research, Proc. from the Intern. Conference, November 25–27, Smolenice, Slovakia, 2009, pp. 19–25.

[5] **L.N. Belyayeva, V.E. Chernyavskaya,** Evidence-based linguistics: methods in cognitive paradigm. Voprosy kognitivnoy lingvistiki [Issues of Cognitive Linguistics], 3 (48) (2016) 77–84.

[6] **S. Bergsma, Q.I. Wang,** Learning noun phrase query segmentation. In Proc. EMNLP- CoNLL, 2007, pp. 819–826.

[7] **V. David, J. Curran,** Adding noun phrase structure to the penn treebank. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Association for Computational Linguistics, Prague, Czech Republic, 2007. pp. 240–247.

[8] **M. Delgado, M.J. Martin-Bautista, D. Sanchez, M.A. Vila,** Mining Text Data: Special Features and Patterns Lecture Notes In Computer Science, Springer-Verlag GmbH, Vol. 2442 (2002) 140–151.

[9] **E. Delpech, B. Daille,** Dealing with lexicon acquired from comparable corpora: validation and exchange, Proceedings, 9th Conference on Terminology and Knowledge Engineering (TKE). Fiontar, Dublin City University, 2010, pp. 229–223.

[10] **R. Feldman, I. Dagan,** Knowledge discovery in textual databases (KDT), Proceedings of the 1st Int. Conference on Knowledge Discovery and Data Mining (KDD-95), AAAI Press, 1995, pp. 112–117.

[11] **E. Heja,** The Role of Parallel Corpora in Bilingual Lexicography. In: N. Calzolari et al. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valetta: European Language Resources Association (ELRA), 2010, pp. 2798–2805.

[12] **A. Lavie, A. Parlikar, V. Ambati,** Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora. In Proc. 2nd SSST, Association for Computational Linguistics, 2008, pp. 87–95.

[13] **M. Malakhovskaya, L. Beliaeva, O. Kamshilova,** Teaching Noun-Phrase Composition in EAP/ESP Context: A Corpus-Assisted Approach to Overcome a Didactic Gap, Journal of Teaching English for Specific and Academic Purposes, 9 (2) (2021) 257–266. DOI: https://doi.org/10.22190/JTESAP2102257M

[14] **N. Reiter, A. Frank,** Identifying Generic Noun Phrases. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, July. Association for Computational Linguistics, 2010, pp. 40–49.

[15] **L. Shen, J. Xu, R. Weischedel,** A new string-to-dependency machine translation algorithm with a target dependency language model. In Proceedings of ACL-08: HLT, 2008, pp. 577–585.

Received 26.04.2021.

СПИСОК ЛИТЕРАТУРЫ

1. **Babych B., Hartley A.** Improving machine translation quality with automatic named entity recognition. In Proceedings of the 7-th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT. Association for Computational Linguistics, 2003. Pp. 1–8.

2. **Baroni M., Zamparelli R.** Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010. Pp. 1183–1193.

3. **Beliaeva L.** Applied Lexicography and Scientific Text Corpora // Transcations on Business and Engineering Intelligent Applications. Galina Setlak, Kassimir Markov (ed.). Rzeshov, Poland: ITHEA, 2014. Pp. 55–63.

4. **Belyaeva L.** Scientific Text Corpora as a Lexicographic Source // SLOVKO 2009. NLP, Corpus Linguistics, Corpus Based Grammar Research, Proc. from the Intern. Conference, November 25–27, Smolenice, Slovakia, 2009. Pp. 19–25.

5. **Беляева Л.Н., Чернявская В.Е.** Доказательная лингвистика: метод в когнитивной парадигме // Вопросы когнитивной лингвистики, 2016, № 3 (48). С. 77–84.

6. **Bergsma S., Wang Q.I.** Learning noun phrase query segmentation. In Proc. EMNLP- CoNLL, 2007. Pp. 819–826.

7. **David V., Curran J.** Adding noun phrase structure to the penn treebank. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Association for Computational Linguistics, Prague, Czech Republic, 2007. Pp. 240–247.

8. **Delgado M., Martin-Bautista M.J., Sanchez D., Vila M.A.** Mining Text Data: Special Features and Patterns // Lecture Notes In Computer Science, Springer-Verlag GmbH, Vol. 2442, 2002. Pp. 140–151.

9. **Delpech E., Daille B.** Dealing with lexicon acquired from comparable corpora: validation and exchange // Proceedings, 9th Conference on Terminology and Knowledge Engineering (TKE). Fiontar, Dublin City University, 2010. Pp. 229–223.

10. **Feldman R., Dagan I.** Knowledge discovery in textual databases (KDT) // Proceedings of the 1st Int. Conference on Knowledge Discovery and Data Mining (KDD-95), AAAI Press, 1995. Pp. 112–117.

11. **Heja E.** The Role of Parallel Corpora in Bilingual Lexicography. In: N. Calzolari et al. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valetta: European Language Resources Association (ELRA), 2010. Pp. 2798–2805.

12. **Lavie A., Parlikar A., Ambati V.** Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora. In Proc. 2nd SSST, Association for Computational Linguistics, 2008. Pp. 87–95.

13. **Malakhovskaya M., Beliaeva L., Kamshilova O.** Teaching Noun-Phrase Composition in EAP/ESP Context: A Corpus-Assisted Approach to Overcome a Didactic Gap // Journal of Teaching English for Specific and Academic Purposes, 2021. Vol. 9. No. 2. Pp. 257–266. DOI: https://doi.org/10.22190/JTESAP2102257M

14. **Reiter N., Frank A.** Identifying Generic Noun Phrases. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, July. Association for Computational Linguistics, 2010. Pp. 40–49.

15. **Shen L., Xu J., Weischedel R.** A new string-to-dependency machine translation algorithm with a target dependency language model. In Proceedings of ACL-08: HLT, 2008. Pp. 577–585.

Статья поступила в редакцию 26.04.2021.

THE AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Beliaeva Larisa N.
Беляева Лариса Николаевна
E-mail: belyaevaln@herzen.spb.ru

Kamshilova Olga N.
Камшилова Ольга Николаевна
E-mail: onkamshilova@gmail.com