Medical vocabulary mining using distributional semantics on Japanese patient blogs

Magnus Ahltorp[1] **Maria Skeppstedt**[2] **Shiho Kitajima**[3] **Rafal Rzepka**[3] **Kenji Araki**[3]
[1]KTH Royal Institute of Technology, Sweden
map@kth.se
[2]Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden
mariask@dsv.su.se
[3]Graduate School of Information Science and Technology, Hokkaido University, Japan
{shiho, rzepka, araki}@ist.hokudai.ac.jp

Abstract

Random indexing has previously been successfully used for medical vocabulary expansion for Germanic languages. In this study, we used this approach to extract medical terms from a Japanese patient blog corpus. The corpus was segmented into semantic units by a semantic role labeller, and different pre-processing and parameter settings were then evaluated. The evaluation showed that similar settings are suitable for Japanese as for previously explored Germanic languages, and that distributional semantics is equally useful for semi-automatic expansion of Japanese medical vocabularies as for medical vocabularies in Germanic languages.

1 Introduction

Distributional semantics models, which are based on word co-occurrence patterns, build on the idea that words frequently occurring in similar contexts often have a similar meaning. There are a number of distributional semantics studies on medical texts written in Germanic languages. These have shown that distributional semantics models, built using large corpora, are useful for semi-automatic medical vocabulary development for Swedish and English (Henriksson et al., 2014; Henriksson et al., 2013) and as features when training English medical named entity recognition systems (Jonnalagadda et al., 2012; Stenetorp et al., 2012; Pyysalo et al., 2014).

Less work has, however, been carried out on medical texts in languages that are grammatically dissimilar from Germanic languages, for instance Japanese. There are previous distributional semantics studies on Japanese, e.g. studies using context in form of noun-verb and noun-noun dependencies (Kazama et al., 2010; Yamada et al., 2009). Research on the adaption of distributional semantics to Japanese corpora using context information in form of several neighbouring words, has, however, not always been successful (Sahlgren et al., 2003), possibly due to the pre-processing and/or parameter choices not being properly adapted. We, therefore, aim to study the usefulness of distributional semantics for constructing medical vocabularies for Japanese, and if pre-processing and parameter settings need to be adapted to Japanese.

2 Background

We have identified three main aspects in which Japanese is different from Germanic languages, and that are relevant when constructing models of distributional semantics: (1) In most distributional semantics studies, the basic semantic unit is formed by the white-space segmented word. White space is, however, not used in Japanese (Kamermans, 2010, p. 17), requiring a different approach for dividing the text into semantic units. (2) A morphologic normalisation in form of a total lemmatisation is often performed on corpora used for distributional semantics. Japanese is, however, highly agglutinative (Tsujimura, 1999, p. 297). Several suffixes can be added to verbs and adjectives, expressing e.g. negation (Kamermans, 2010, p. 54) or desire (Kamermans, 2010, p. 111). Full lemmatisation might, therefore, result in severe loss of information. (3) Distributional semantics studies on Germanic languages have shown that taking information from a small context window of co-occurring words into account (typically 1–2 preceding and following words) is most suitable when building models for similarity between words (Sahlgren et al., 2008). This does not necessarily have to be the case for Japanese. This is partly due to that the basic word order is different (SOV), but also due to that the word order is relatively free and the function of a word (e.g.

whether it is a topic, subject or object) instead is indicated by case particles (Kamermans, 2010, pp. 35-38). Therefore, another context window size might be more appropriate for Japanese. Also the stopword filtering, often used for e.g. English vocabulary extraction (Sahlgren et al., 2008), might have to be adapted to Japanese, possibly retaining the frequently occurring case particles.

3 Method

To address that white space is not used in Japanese, a number of steps were applied for segmenting the corpus into semantic units. A basic pre-processing was first carried out by removing smileys and sentences solely containing Latin characters, as well as normalising the syllable writing characters *hiragana* and *katakana* by transforming half width forms into the corresponding full width form. Then the corpus was segmented into semantic units by: (a) Applying the dependency parser CaboCha on the corpus (CaboCha, 2012) (b) Applying the semantic role labeller ASA (ASA, 2013) on the parsed corpus.

Three different pre-processing versions were applied to the corpus, to study effects of the agglutinative nature of Japanese and of word order differences. In the *Standard* version, parts of the information contained in the suffixes, which potentially has a large impact on the semantics of the surrounding semantic units, was retained. This included polarity (negation or affirmation), grammatical mood and voice, while e.g. formality level and tense were excluded. In this Standard version, stopword filtering was also carried out by removing all semantic units not classified by CaboCha as either a verb (not including helper verbs or copula), an adjective (including adverbial derivations of adjectives) or a noun (including pronouns). In the first alternative pre-processing version (*Total lemmatisation*), no information from the suffixes was retained, but the corpus was instead completely lemmatised. In the second alternative version (*With case particles*), the Standard pre-processing version was modified by retaining case particles, to study if this could compensate for the relatively free word order of Japanese. Multiple adjacent case particles were grouped into one semantic unit.

Effects of word order differences were also studied by varying the context window size. Context window sizes of 1+1, 2+2, 4+4 and 8+8 surrounding semantic units[1] were evaluated for each one of the three pre-processing versions. This resulted in a total of 12 created semantic spaces.

The corpus, from which the distributional semantics models were built, was a Japanese blog corpus from the TOBYO site, which collects blogs written by patients and/or their relatives (Kitajima et al., 2013). After the first normalising step, the corpus contained 270 million characters and the *standard* pre-processing version contained 50 million semantic units (2.5 million unique).

3.1 Random Indexing

For constructing the distributional semantics models, *random indexing* was chosen, because of its computational efficiency on large corpora (Sahlgren and Karlgren, 2009). In random indexing, which was introduced by Kanerva et al. (2000), each semantic unit of segmented text is assigned a representation in form of an *index vector*, all of the same relatively small dimension. The index vectors have most elements set to 0, except a few, randomly selected elements, who are either given the value +1 or –1. Each semantic unit is also assigned a *context vector* of the same small dimension, and with all elements initially set to 0. The context vectors are then updated to represent the context of the semantic unit, by, for each occurrence of the unit in the corpus, adding the index vectors of the surrounding semantic units. The resulting context vectors form an approximation of a standard co-occurrence matrix, and similarity between semantic units can be measured by e.g. the cosine of the angle between the context vectors.

The configuration was generally based on results by Sahlgren et al. (2008), and non-weighted *direction vectors* (of dimension 40,000) were therefore used. This encodes whether a surrounding semantic unit occurs to the right or left of the target unit, but does not encode its distance to the target. 400 non-zero elements were used in the index vectors.

3.2 Evaluation

The semantic spaces were evaluated for their ability to extract terms belonging to a certain semantic category, given a number of seed terms of this category. Three semantic categories, highly relevant for patient blogs, were used: *Medical Finding*, *Pharmaceutical Drug* and *Body Part*. As seed

[1] Not letting the context cross a sentence boundary.

terms and as evaluation data, terms from Japanese MeSH were used; terms classified under the node *Diseases (C)* for Medical Finding, terms classified under the node *Chemicals and Drugs (D)* for Pharmaceutical Drugs and terms classified under the node *Anatomy (A)* for *Body part*[2]. Pharmaceutical brand names available at the TOBYO site (Tobyo, 2013), as well as terms from a language education web page listing body parts in Japanese (Chonmage Eigojuku, 2013), were also included to decrease the difference in number of terms between the category Medical Finding and the other two categories. Terms occurring at least 50 times in the segmented corpus as a semantic unit in the context of at least one other semantic unit, were included in the set of used terms. Terms in existing vocabularies that were segmented into several semantic units were, therefore, excluded, as were infrequent terms, due to the weak statistical foundation for the position of their context vectors. This resulted in 331 terms for Medical Finding, 278 for Pharmaceutical Drug and 214 for Body Part.

The sets of terms were divided into two equally large groups, group A and group B. The terms in group A were first used as seed terms and the terms in group B as reference standard. Thereafter, the reverse setup was applied; terms in group B were used as seed terms and terms in group A as reference standard. The seed terms represent terms that, in a real-world scenario, would be included in an existing but incomplete vocabulary, while the reference standard represents terms that should be suggested by an optimal semantic space for inclusion in the vocabulary. As the two setups were treated as two separate evaluations, using the seed terms from one setup as reference standard in the other does not bias the results in any way. The final results were calculated by averaging the results from the two separate evaluations.

The sets of seed terms were used for ranking the other semantic units in the created random indexing spaces according to their similarity to seed terms of each semantic category. The ranking was obtained with the method used by Skeppstedt et al. (2013), in which a *summed similarity* to a semantic category was calculated for every semantic unit in the random indexing space. This was calculated by summing the cosine of the angle ($\theta_{\bar{u},\bar{s}}$) between the context vector of the semantic unit ($\bar{u}$) and the context vector of each term ($\bar{s}$) in the set of seed terms (S) of the semantic category in question.

$$\textit{summed similarity}(\bar{u}) = \sum_{\bar{s} \in S} \cos(\theta_{\bar{u},\bar{s}})$$

All semantic units (apart from the seed terms) occurring at least 50 times in the corpus were ranked according to its summed similarity, resulting in one ranked list for each of the three semantic categories. For each of the 12 created random indexing spaces, and each A/B-group setup, one such triplet of ranked lists was produced and evaluated.

Recall for retrieving the terms in the reference standard was calculated by measuring whether the terms in the reference standard were included among the top $i * n$ highest ranked semantic units, where n is the number of terms in the reference standard for a given category and setup and where i was given integer values between 1 and 10.

For calculating precision for retrieving terms of the correct semantic category, regardless of if a retrieved term was included in existing vocabularies, a small manual evaluation was also carried out. The top 150 retrieved semantic units in the ranked lists for Medical Finding and Body Part were manually classified according to semantic category. Semantic units were presented to the annotators (one with basic level of Japanese and one with intermediate level of Japanese) along with an English translation generated using JMDict (JMDict, 2013), and without revealing which semantic space had produced the candidate. For borderline cases, the semantic category of the English version of SNOMED CT (IHTSDO, 2008) was used as a reference. Semantic units that the annotators were not able to classify (e.g. since no translation was provided by JMDict), as well as semantic units for which the two annotators disagreed on the classification, were classified by a third annotator, a native Japanese speaker.

4 Result and Discussion

Results for evaluated semantic spaces are shown in Figure 1. *Standard* pre-processing and *total lemmatisation* produced similar results for all three semantic categories, which shows that retaining some of the inflected information does not contribute to the term extraction.

Retaining case particles for Medical Finding and Pharmaceutical Drug produced low results, regardless of window size. Removing case particles and using a window size of 1+1 was instead

[2]Except terms under the sub-nodes *Plant Structures (A18)*, *Fungal Structures (A19)*, *Bacterial Structures (A20)* and *Viral Structures (A21)*.

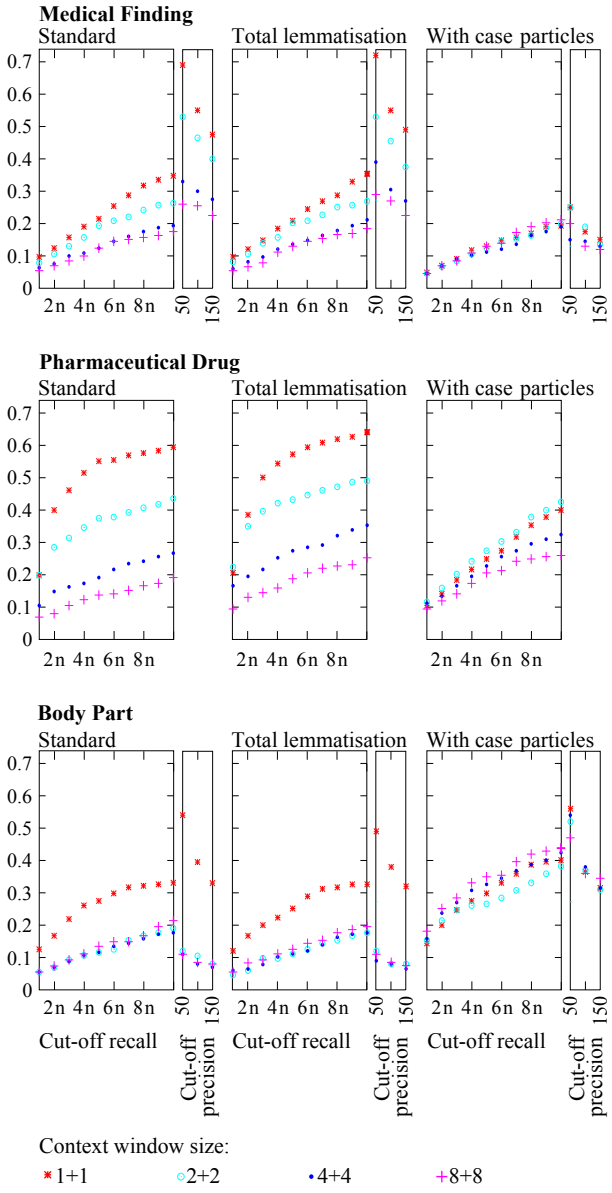

Figure 1: *Recall* (to the left) was measured for top n to top $10n$ candidate terms (n = the number of terms in the reference standard). *Precision* (to the right for Medical Finding and Body Part) was measured for top 50, 100 and 150 candidate terms. Existing vocabularies were used as reference standard for recall, while the reference standard for precision was extended by manual classification.

clearly the most successful configuration for those two categories. For Body Part, on the other hand, the best recall results were achieved when retaining case particles. Using the standard stopword filtering with a window size of 1+1 produced, however, only slightly lower recall, and precision results were similar for the two approaches.

The manually classified reference standard, against which precision was evaluated, had an inter-annotator agreement of 95%, excluding the class Unknown, which the annotators used in 9.3% (basic Japanese) and 0.4% (intermediate Japanese) of the cases. Despite being evaluated against a different reference standard, precision results followed the same pattern as those for recall. This shows a potential for relying on the efficient method of using fully automatic recall evaluation against existing vocabularies for determining what parameters to use. Precision, however, is needed to compare between combinations of language, corpus and vocabulary, since the completeness of used vocabularies have a large impact on recall.

For extraction of Swedish Medical Findings with the summed similarity ranking (Skeppstedt et al., 2013) slightly better precision was reported (0.80, top 50, 0.68, top 100 and ∼0.58 top 150, compared to 0.72, 0.55 and 0.49 achieved here). This difference could be due to the additional complexity associated with segmenting Japanese text, but could also be explained by other non-language factors, e.g. different corpora types and suitability of the seed terms to the corpus type.

5 Conclusion

Results showed that parameters suitable for previously explored Germanic languages (a small window size, stopword filtering of function/frequent words and total lemmatisation) are suitable also for Japanese. An exception was, however, the category Body Part, for which slightly better results were achieved when case particles were retained.

The achieved precision for Medical Finding is only slightly lower than for a similar study on Swedish (Skeppstedt et al., 2013), showing that distributional semantics has the same potential to be useful for semi-automatic expansion of Japanese medical vocabularies as for the previously explored languages.

The consistency between recall and precision increases our confidence in the results. The confidence could, however, be further increased by using several resamplings of the division into seed terms and evaluation terms. Future work also includes application of the evaluated techniques for Japanese named entity recognition, as well as for semi-automatic expansion of Japanese medical vocabularies to also include terms typical to the language used in patient blogs.

Acknowledgments

We would like to thank the three anonymous reviewers for their many good comments.

References

ASA. 2013. Semantic role tagger for Japanese language. http://cl.cs.okayama-u.ac.jp/study/project/asa.

CaboCha. 2012. Cabocha: Yet another Japanese dependency structure analyzer. https://code.google.com/p/cabocha/.

Chonmage Eigojuku. 2013. Chonmage Eigojuku, English names of body parts. http://mage8.com/tango/tango3.html.

Aron Henriksson, Mike Conway, Martin Duneld, and Wendy W. Chapman. 2013. Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records. In *Proceedings of the Annual Symposium of the American Medical Informatics Association (AMIA 2013)*, Washington DC, USA.

Aron Henriksson, Hans Moen, Maria Skeppstedt, Vidas Daudaravičius, and Martin Duneld. 2014. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *J Biomed Semantics*, 5(1):6.

IHTSDO. 2008. SNOMED Clinical Terms User Guide, July 2008 International Release, International Health Terminology Standards Development Organisation. http://www.ihtsdo.org. International Release 2013-01-31.

JMDict. 2013. The JMDict project. http://www.edrdg.org/jmdict/j_jmdict.html.

Siddhartha Jonnalagadda, Trevor Cohen, Stephen Wu, and Graciela Gonzalez. 2012. Enhancing clinical concept extraction with distributional semantics. *J Biomed Inform*, 45(1):129–40, Feb.

Michiel Kamermans. 2010. *An Introduction to Japanese Syntax, Grammar & Language*. Sjgr Publishing.

Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In L. R. Gleitman and A. K. Joshi, editors, *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, Mahwah, NJ.

Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, and Kentaro Torisawa. 2010. A bayesian method for robust estimation of distributional similarities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 247–256, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shiho Kitajima, Rafal Rzepka, and Kenji Araki. 2013. Performance improvement of drug effects extraction system from Japanese blogs. In *Proceedings of 2013 IEEE Seventh International Conference on Semantic Computing*, pages 383–386, Irvine, USA.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2014. Distributional semantics resources for biomedical text processing. In *Proceedings of Languages in Biology and Medicine*.

Magnus Sahlgren and Jussi Karlgren. 2009. Terminology mining in social media. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09.

Magnus Sahlgren, Preben Hansen, and Jussi Karlgren. 2003. English-Japanese cross-lingual query expansion using random indexing of aligned bilingual text data. In *Proceedings of the Third NTCIR Workshop*.

Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a means to encode order in word space. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 1300–1305.

Maria Skeppstedt, Magnus Ahltorp, and Aron Henriksson. 2013. Vocabulary expansion by semantic extraction of medical terms. In *Proceeding of Languages in Biology and Medicine (LBM 2013)*.

Pontus Stenetorp, Hubert Soyer, Sampo Pyysalo, Sophia Ananiadou, and Takashi Chikayama. 2012. Size (and domain) matters: Evaluating semantic word space representations for biomedical text. In *Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine*.

Tobyo. 2013. Tobyo pharmaceutical drugs. http://www.tobyo.jp/reference/1-2-1.php?KEY_INDEX=2&key=1.

Natsuko Tsujimura. 1999. *The handbook of Japanese linguistics*. Blackwell Publishers, Malden, Mass.

Ichiro Yamada, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, Masaki Murata, Stijn De Saeger, Francis Bond, and Asuka Sumida. 2009. Hypernym discovery based on distributional similarity and hierarchical structures. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 929–937, Stroudsburg, PA, USA. Association for Computational Linguistics.