CroVeWA: Crosslingual Vector-Based Writing Assistance

Hubert Soyer[1]*, **Goran Topić**[1], **Pontus Stenetorp**[2]†**and Akiko Aizawa**[1]
[1] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
[2] University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
{soyer,goran_topic,aizawa}@nii.ac.jp pontus@stenetorp.se

Abstract

We present an interactive web-based writing assistance system that is based on recent advances in crosslingual compositional distributed semantics. Given queries in Japanese or English, our system can retrieve semantically related sentences from high quality English corpora. By employing crosslingually constrained vector space models to represent phrases, our system naturally sidesteps several difficulties that would arise from direct word-to-text matching, and is able to provide novel functionality like the visualization of semantic relationships between phrases interlingually and intralingually.

1 Introduction

Writing high quality texts in a foreign language requires years of study and a deep comprehension of the language. With a society that is becoming more and more international, the ability to express ideas in English has become the basis of fruitful communication and collaboration.

In this work, we propose a tool to provide non-native speakers of English with help in their translation or writing process. Instead of relying on manually created dictionaries, many existing tools leverage parallel bilingual corpora, using a concordancer to provide translation suggestions together with their contexts. Notable examples relevant to this demonstration are `linguee.com` and `tradooit.com`. Given a word or a phrase in a foreign language, these systems present example sentences containing the query in the source language as well as the target language, showing the correct usage of the word/phrase, and at the same time providing translation candidates.

Many applications rely on direct word-to-text matching and are therefore prone to missing semantically similar contexts that, although similar and relevant, do not share any words with the query. Instead of matching words directly, we propose a system that employs crosslingually constrained vector representations (embeddings) of words and phrases to retrieve English sentences that are similar to a given phrase or word in a different language (query). These vector representations not only allow for efficient crosslingual lookups in databases consisting of millions of sentences, but can also be employed to visualize intralingual and interlingual semantic relationships between phrases.

2 Related Work

Various types of neural network models have been proposed to induce distributed word representations and leveraging these word embeddings as features has proven viable in achieving state-of-the-art results for a variety of tasks (Baroni et al., 2014; Collobert and Weston, 2008).

Recently, methods that attempt to compose embeddings not only of words but of whole phrases (Le and Mikolov, 2014; Socher et al., 2011) have enabled vector representations to be applied for tasks that are defined over phrases, sentences, or even documents. The most relevant work for this paper are recent approaches that allow for the induction of

*Currently at Google DeepMind.
†Currently at University College London.

Proceedings of NAACL-HLT 2015, pages 91–95,
Denver, Colorado, May 31 – June 5, 2015.

word and phrase embeddings not only from monolingual text but using bilingual resources to constrain vector representations crosslingually. (Soyer et al., 2015; Hermann and Blunsom, 2014; Cho et al., 2014; Chandar A P et al., 2014). Embeddings learned using these methods not only possess meaningful properties within a language, but also interlingually.

3 Crosslingual Vector-Based Writing Assistance (CroVeWA)

Our system harnesses crosslingually constrained word and phrase representations to retrieve and visualize sentences related to given queries, using distances in the word/phrase vector space as a measure of semantic relatedness. Currently, our system supports the lookup of Japanese and English queries in English text.

Our system encourages refining retrieved results and viewing relations in different contexts by supporting multiple queries. All queries and their corresponding results are visualized together to aid a better understanding of their relationships. To illustrate the differences to phrase vector-based sentence retrieval, we also offer a retrieval option based on direct word-to-text matching using the EDICT Japanese-English dictionary (Breen, 2004) and Apache Lucene[1] for sentence retrieval.

To the best of our knowledge, our system is the first to provide writing assistance using vector representations of words and phrases.

3.1 Inducing Crosslingually Constrained Word Representations

We employ the approach presented in Soyer et al. (2015) to learn bilingually constrained representations of Japanese and English words. The method draws from sentence-parallel bilingual text to constrain word vectors crosslingually, handles text on a phrase level ensuring the compositionality of the induced word embeddings, and is agnostic to how phrase representations are assembled from word representations. In addition, unlike previously proposed models, the model can draw not only from bilingual sentence aligned data but also from arbitrary monolingual data in either language. Figure 1 depicts an overview over the method.

The method optimizes the vectors that represent each word in subject to a bilingual and a monolingual objective. These objectives operate on a phrase level, where each phrase is represented by a single vector. Composing a single vector of a given phrase means looking up the word vector for each word in a lookup table shared among all sentences of the phrase-language, and applying a composition function to collapse all word vectors of a phrase into a single phrase vector. The composition function used in this work is the arithmetic mean.

The *bilingual objective* ensures that vectors of Japanese sentences are close to the vectors of their English translations present in the sentence-parallel corpus. It minimizes the squared euclidean distance between the sentence vector of a Japanese sentence and the vector of its English translation. With the arithmetic mean as the sentence composition function, this notion of translational closeness is directly propagated back into the embeddings of the individual words that appear in each sentence. If a Japanese and an English word consistently co-occur in the translation pairs of the sentence-parallel corpus, their vectors will be moved close to each other, capturing that they are likely to be related in meaning.

The *monolingual objective* exploits the insight that sub-phrases generally tend to be closer in meaning to the phrases they are contained in, than to most other arbitrary phrases. It punishes a large euclidean distance between the vector representation of a phrase and its sub-phrase, and at the same time rewards a large distance between the vector of the phrase and the embedding of another phrase chosen at random.

Both the monolingual objective and the bilingual objective are combined to leverage monolingual and bilingual resources at the same time. Using the arithmetic mean to compose phrase vectors discards word-order as well as sentence-length information, and allows our system to handle even single words or ungrammatical sequences of words.

Currently we use Japanese and English resources to learn word embeddings, but plan to add more languages in the future. The bilingual sentence-parallel

[1] https://lucene.apache.org/core/

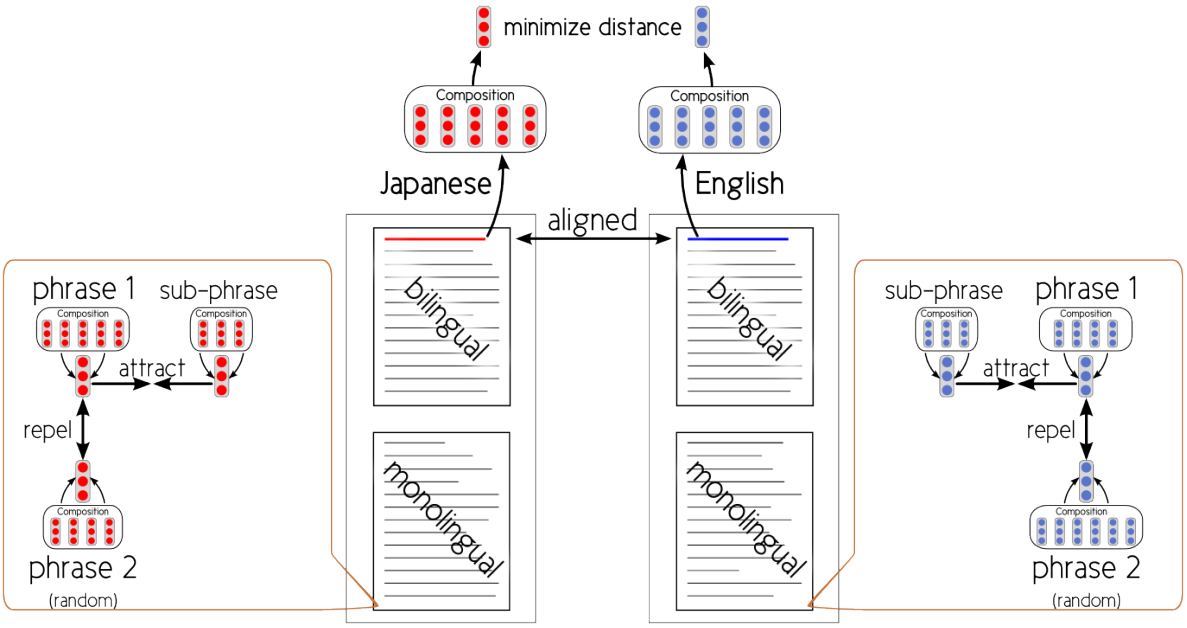

Figure 1: Overview of the method that was used to induce crosslingually constrained word representations. The method can draw from bilingual sentence-parallel data as well as monolingual data.

resource used is the ASPEC corpus[2], which features sentence-aligned text from scientific paper abstracts. For monolingual data, we use subsets of the Japanese and English Wikipedia.

3.2 Finding Related English Sentences for a Japanese Query

Inducing crosslingually constrained word representations leaves us with two sets of vectors, one corresponding to Japanese words and one to English words. Given a query in **Japanese**, we look up the vectors for each individual query word, compose them into a single query vector and find the nearest neighbors in a set of pre-computed vectors of **English** sentences. Since the word and phrase vectors are crosslingually constrained, we expect the retrieved English nearest neighbors to be semantically related to the Japanese query. In contrast to conventional word matching techniques, our vector-based approach does not require Japanese translations of the English sentences we consider during the search, nor does it require a Japanese-English dictionary.

Another difference to word matching techniques follows from the way word vectors are arranged within the same language. Generally, words that appear in similar contexts will be placed close to each other in the vector space, and so the difference between choosing a word over a closely related neighbor will be relatively small when composing a phrase vector. Interchangeability of synonyms or semantically similar words is therefore automatically supported as a property of the word representations, and the system can retrieve sentences similar in meaning regardless of the exact choice of words.

Following Mikolov et al. (2013) we use the cosine similarity as a measure of similarity between embeddings. For nearest neighbor retrieval we employ the FLANN Python module (Muja and Lowe, 2009) which exploits the clustered nature of vector representations to efficiently find an approximate set of nearest neighbors.

3.3 Visualization

In contrast to direct word matching, vector-representation-based matching retrieves not only a list of related sentences, but also a semantic vector space position for each query and result. In order to visualize the high-dimensional output vectors of the search we reduce their dimensionality to two.

Generally, reducing dimensionality involves dis-

[2] http://orchid.kuee.kyoto-u.ac.jp/ASPEC/

carding information. Commonly employed methods for this task such as, Principal Component Analysis or t-SNE (Van der Maaten and Hinton, 2008), failed to provide satisfactory results for our purposes. Instead, we apply a novel variant of multidimensional scaling (Kruskal, 1964) where we prioritize the preservation of query-to-result distances over the preservation of result-to-result distances. This yields a visually more interpretable output, with queries being the points of orientation.

An interactive plot of the resulting 2D points, illustrates the relationships between the different sentences, puts the retrieved results into context and aids the user's understanding of the meaning and relatedness of sentences. Being able to visualize these relationships is another aspect that sets our system apart from previously proposed word-to-text matching approaches.

3.4 Demonstration System

We will guide the audience through the features of our application using a set of example queries that highlight the merits and drawbacks of vector-based crosslingual sentence retrieval. Based on these query examples we will introduce novel functionality, such as our system's visualization of semantic relationships or its feature for query auto-generation. The interactive nature of our tool allows us to incorporate requests and comments into the demonstration, helping to clarify questions and to explain properties of our system.

Our system is built as a web application and therefore only requires the user to have a modern browser and an Internet connection. Figure 2 shows a screenshot of the user interface, which consists of the query input bar at the top of the screen, a result list on the left and a visualization panel on the right. For clarity, we have annotated the screenshot: annotations with white background show results and their positions in the visualization, while the ones with red background provide translations of the Japanese queries.

In the query input bar users can customize the search through a variety of options. Via the query input field a user can submit Japanese queries which can be a single word, a phrase or any sequence of words. Pushing the *Auto-Generate* button will split the entered text into semantically related groups of words and submit these groups as separate queries to visualize the relatedness of different parts of the entered text. Since not every potential user might be familiar with Japanese we provide an *English simulation mode* to input English queries and retrieve English results. We refer to this mode as *simulation* because the lookup from English to English is not crosslingual. For comparison, and as an extension to the vector-based sentence retrieval, we also provide a dictionary-based word-to-text matching search mode using the Japanese-English EDICT dictionary. Clicking the *Samples* button invokes a dialog that presents example queries to choose from. We currently provide three corpora to search, where each corpus covers a different domain. The ASPEC corpus consists of Japanese and English scientific paper abstracts related to natural sciences and engineering, the Wikipedia corpus comprises 10 million randomly selected sentences from the English Wikipedia, and the PubMed corpus features 5 million sentences from the PubMed Central collection of medical paper abstracts.

Queries make up the first entries in the result list, with fully colored backgrounds. In the visualization panel queries are represented by larger points. If two or more queries have been submitted we additionally provide a *Query Average* to retrieve results that are related to all submitted queries. Every result entry that follows the queries is colored according to the closest query from those that retrieved it. The fill level of the background of each result item indicates its similarity to the *Query Average*. Hovering the cursor over a list entry will highlight its corresponding point and vice versa. Clicking on a list entry will auto-generate queries from its text, clustering related words together to provide a visualization of the different topics the text consists of.

The annotations in Figure 2 illustrate how the system's visualization can aid a user in understanding the meaning of a sentence. The distance between a result and a query indicates their semantic closeness. Sentences located close to a query point are strongly related to the query (*canal* and *river* or *subway station* and *railway station*), phrases in between the queries feature aspects of both submitted queries (*flooding*, *subway* and *city*).

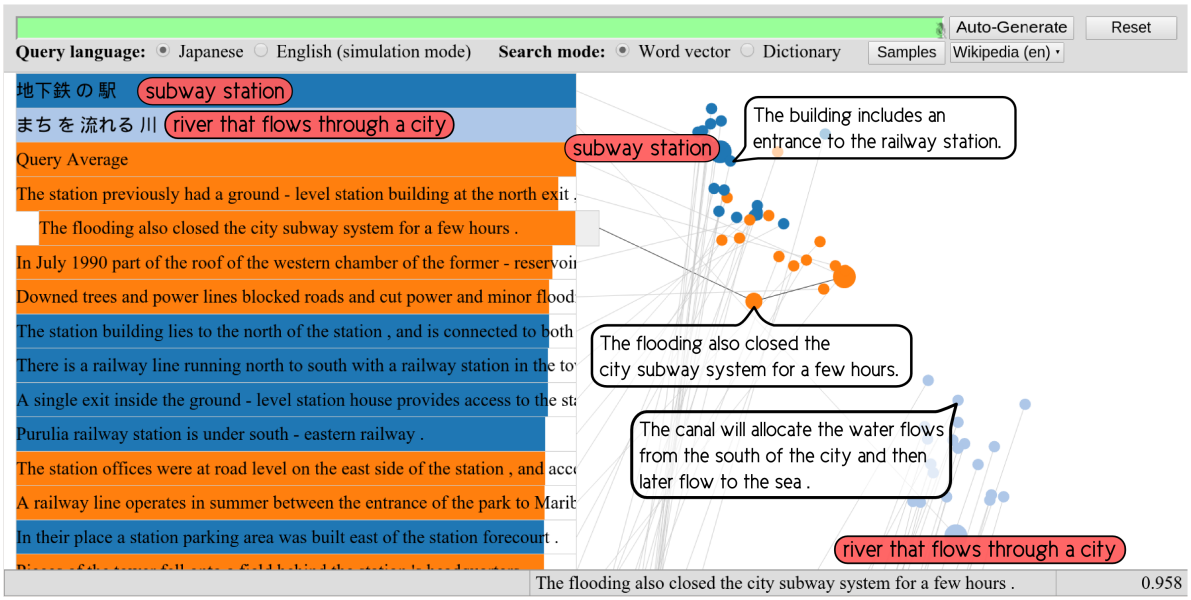

Figure 2: Annotated screenshot of CroVeWA.

Acknowledgements

This work was supported by the Data Centric Science Research Commons Project at the Research Organization of Information and Systems and by the Japan Society for the Promotion of Science KAKENHI Grant Number 13F03041.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd ACL*, pages 238–247.

James Breen. 2004. JMDict: a Japanese-multilingual dictionary. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, pages 71–79. ACL.

Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1853–1861. Curran Associates, Inc.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of EMNLP 2014*, pages 1724–1734.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th ICML*, pages 160–167. ACM.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd ACL*, pages 58–68.

Joseph B Kruskal. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of The 31st ICML*, pages 1188–1196.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Marius Muja and David G. Lowe. 2009. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Application VISSAPP'09)*, pages 331–340. INSTICC Press.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of EMNLP*, pages 151–161. Association for Computational Linguistics.

Hubert Soyer, Pontus Stenetorp, and Aizawa Akiko. 2015. Leveraging monolingual data for crosslingual compositional word representations. In *Proceedings of ICLR*. to appear.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR*, 9(2579-2605):85.