A resource of sentence analogies on the level form extracted from corpora in various languages

Rashel Fam Yves Lepage
早稲田大学大学院 情報生産システム研究科
fam.rashel@fuji.waseda.jp yves.lepage@waseda.jp

Abstract

Word analogy datasets are commonly used to assess the quality of word embeddings. As the NLP tasks are going more and more towards sentences and beyond, vector representation of these units is becoming more and more vital to the performance of the system. However, there are not so many datasets available for sentence analogy. In this paper, we release a resource of analogies between sentences extracted from two corpora: Tatoeba and Multi30K. The analogies are extracted in various European languages.

1 Introduction

I like coffee : *I like tea* :: *I like hot coffee* : *I like hot tea*
⇔
I like hot coffee : *I like hot tea* :: *I like coffee* : *I like tea*
⇔
I like coffee : *I like hot coffee* :: *I like tea* : *I like hot tea*

Figure 1 Examples of analogy on sentences with their equivalent analogies derived from properties of analogy mentioned in Section 1: symmetry of conformity and exchange of the means.

Analogy is a relationship between four objects: A, B, C and D where A is to B as C is to D. It is noted as $A : B :: C : D$. As our work relates to strings, A, B, C and D are all strings (sequence of characters). This notation means that the ratio between A and B is similar to the ratio of C and D. In another way, an analogy is a conformity of ratios between the four strings, as shown in Formula (1). Figure 1 gives examples of analogy between sentences.

$$A : B :: C : D \overset{\Delta}{\iff} \begin{cases} A : B = C : D \\ A : C = B : D \end{cases} \quad (1)$$

In this paper, we adopt the definition of formal analogies between strings of symbols as found in [1, 2, 3].

2 Number of analogies in a text and analogical density

We address the theoretical problem of counting the total number of analogies in a given text. The following section will introduce two main metrics used in this work.

2.1 Analogical density

The analogical density (D_{nlg}) of a corpus is defined as the ratio of the total number of analogies contained in the corpus (N_{nlg}) against the total number of permutations of four objects that can be constructed by the number of sentences (N_s).

$$D_{nlg} = \frac{N_{nlg}}{\frac{1}{8} \times N_s^4} = 8 \times \frac{N_{nlg}}{N_s^4} \quad (2)$$

The factor $1/8$ in the denominator comes from the fact that there exist 8 equivalent forms of an analogy due to two main properties of analogy:

- symmetry of conformity: $A : B :: C : D \Leftrightarrow C : D :: A : B$, and
- exchange of the means: $A : B :: C : D \Leftrightarrow A : C :: B : D$.

2.2 Proportion of sentences appearing in analogy

We count the number of sentences appearing in at least one analogy (N_{s_nlg}) and take the ratio with the total number of sentences in the corpus (N_s) to get the proportion of sentences appearing in at least one analogy (P).

$$P = \frac{N_{s\_nlg}}{N_s} \quad (3)$$

3 Original corpora

We consider two corpora to use in this work, Tatoeba and Multi30K. These two corpora are available on the web and

already heavily used by the natural language processing community.

- **Tatoeba**[1]: is a collection of sentences that are translations provided through collaborative works online (crowd-sourcing). It covers hundreds of languages. However, the amount of data between languages is not balanced because it also depends on the number of members who are native speakers of that language. Sentences contained in Tatoeba corpus are usually short. These sentences are mostly about daily life conversations.
- **Multi30K**[2] [4, 5, 6]: is a collection of image descriptions (captions) provided in several languages. This dataset is mainly used for multilingual image description and multimodal machine translation tasks. It is an extension of Flickr30K [7] and more data is added from time to time, for example, COCO dataset[3].

Table 1 provides the statistics on Tatoeba and Multi30K. As an overview, Multi30K has two times number of tokens in a sentence in comparison to Tatoeba, These two corpora can be characterised based on the diversity of the context of the sentence it contains. Multi30K is a corpus with diverse contexts. In comparison to that, sentences contained in Tatoebaare less diverse. Tatoeba is mostly about daily life conversation. We expect that corpus with less diversity of context will share words between sentences more often. Thus, it will have more analogies and higher analogical density.

Let us now compare the statistics between languages. English has the lowest number of types. Finnish, Polish and Czech always have the highest number of types for around two times higher than English across the corpora. We can observe that language with poor morphology has fewer of types and hapaxes. On the contrary, languages with high morphological richness have less number of tokens due to richer vocabulary. These languages also tend to have longer words (in characters). One can easily understand that with richer morphological features we will have a higher vocabulary size. The consequence of this is that the words will be longer. We also observe that a higher number of types means lesser words to repeat (higher Type-Token-Ratio). Thus, the number of tokens will be lower.

However, we also see that there are some interesting exceptions. In this case, French and German. French has a higher number of tokens than English despite having a higher vocabulary size. The high number variety of function words (propositions, articles, etc.) in French is probably one of the explanations of this phenomenon. As for German, it has a pretty high average length of type in comparison to other languages. This is maybe caused by words in German are originally already longer. German is known to glue several words as a compound word.

1) tatoeba.org
2) github.com/multi30k/dataset
3) cocodataset.org

4 Newly created sentence analogy dataset

4.1 Extraction of analogies

To extract analogies from a corpus, we rely on an already existing tool described in [8]. It relies on the equality of ratios as the definition of analogy.

Each sentence in the corpus is represented as a vector shown in Formula (4). We use the notation $|S|_c$ which stands for the number of occurrences of token c in string S. The number of dimensions of the vector is the size of the alphabet or the vocabulary, depends on the tokenisation scheme (See Section 4.2).

$$A \triangleq \begin{pmatrix} |A|_{t_1} \\ |A|_{t_2} \\ \vdots \\ |A|_{t_N} \end{pmatrix} \tag{4}$$

The conformity between ratios of strings is defined as the equivalent between the two vectors of ratios. See Formula (5).

$$A : B :: C : D \quad \overset{\Delta}{\Longleftrightarrow} \quad \begin{cases} A : B & = & C : D \\ A : C & = & B : D \end{cases} \tag{5}$$

Pairs of strings representing the same ratio can be grouped as an analogical cluster. Please refer to Formula (6). Notice that the order of string pairs has no importance.

$$\begin{array}{l} A_1 : B_1 \\ A_2 : B_2 \\ \quad \vdots \\ A_n : B_n \end{array} \quad \overset{\Delta}{\Longleftrightarrow} \quad \begin{array}{r} \forall (i, j) \in \{1, \ldots, n\}^2, \\ A_i : B_i :: A_j : B_j \end{array} \tag{6}$$

Table 1 Statistics on Tatoeba and Multi30K.

	-	en	fr	de	cs	pl	fi
Tatoeba	# of lines	7,964					
	# of tokens	51,279	54,430	50,375	-	41,892	39,907
	# of types	4,152	5,740	5,639	-	7,796	8,634
	Avg. tokens per line	6.44±2.80	6.83±3.20	6.33±2.85	-	5.26±2.44	5.01±2.10
	Avg. token length	3.34±2.14	3.69±2.52	4.04±2.59	-	4.26±2.89	4.75±3.15
	Avg. type length	6.32±2.27	7.12±2.49	7.55±3.01	-	7.28±2.44	8.09±2.86
	Type-Token-Ratio	0.08	0.11	0.11	-	0.19	0.22
	Hapax size (%)	48.80	56.17	55.68	-	62.11	66.50
Multi30K	# of lines	30,014					
	# of tokens	392,978	471,352	374,490	308,367	-	-
	# of types	10,373	11,376	19,112	22,787	-	-
	Avg. tokens per line	13.09±4.10	15.70±5.91	12.48±4.23	10.27±3.60	-	-
	Avg. token length	3.85±2.40	3.93±2.47	4.86±2.97	4.34±2.71	-	-
	Avg. type length	6.92±2.41	7.41±2.42	9.91±3.91	7.52±2.40	-	-
	Type-Token-Ratio	0.03	0.02	0.05	0.07		
	Hapax size (%)	41.94	42.15	58.05	53.50	-	-

Table 2 Example sentences (lowercased and tokenised) randomly chosen from corpora used in the experiment. Sentences contained in the same corpus are translations of each other in the other languages.

		Example sentences
Tatoeba	en	*the store is closing at 7 .*
	fr	*le magasin ferme à 7 heures .*
	de	*der laden schließt um sieben .*
	pl	*sklep jest zamkniety od 19 .*
	fi	*kauppa menee kiinni kello seitsemän .*
Multi30K	en	*a boy in white plays baseball .*
	fr	*un garçon en blanc joue au baseball .*
	de	*ein weiß gekleideter junge spielt baseball .*
	cs	*chlapec v bílém hraje baseball .*

4.2 Tokenisation schemes

The sentence is tokenised using different tokenisation schemes: character, sub-word and word. For sub-word[4], we use two popular sub-word models: unigram language model (unigram) [9] and byte-pair-encoding (BPE) [10]. The delimiter used to separate tokens is the space. Underscores denote spaces in the original sentence. The vocabulary size used here for unigram and BPE is 1,000 (1k).

4.3 Statistics on the newly created dataset

Each of the corpora is tokenised using four different tokenisation schemes: character, BPE. unigram and word. On top of that, we performed masking with both methods: the least frequent and most frequent. Ablation experiments are carried out on all corpora in six languages depending on the language availability of the corpus.

In this paper, we decided to carry out the experiment on both Tatoeba and Multi30K as these corpora have a different range on both formal and semantic levels. On the formal level, sentences in Tatoeba are short and similar to one another. Multi30K contains more diverse and longer sentences. On the level of semantics, as mentioned in Section 3, Tatoeba contains sentences that focus on the theme of daily conversation. Multi30K, which contains image captions, has a wider range of topics.

Figure 2 (top-left) shows the number of analogical clusters extracted from the corpora with various tokenisations in English. Tatoeba has the highest number of clusters. This meets our hypothesis. It is also reflected in the num-

4) github.com/google/sentencepiece

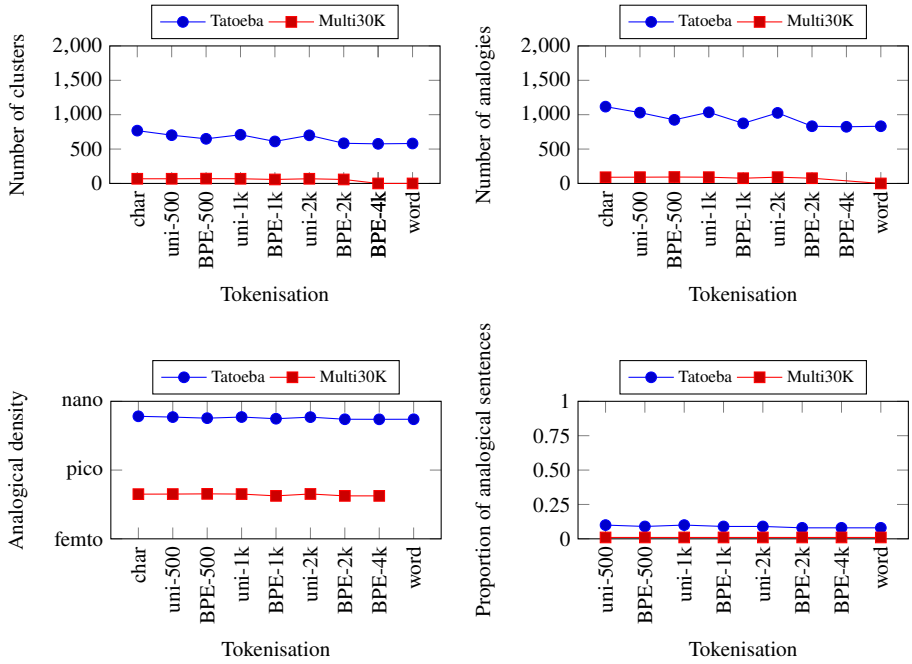

Figure 2 Number of clusters (*top-left*) and analogies (*top-right*) extracted from the corpora in English. Below that, Analogical density (*bottom-left*) and the proportion of sentences appear in analogy (*bottom-right*) for the corpora in English. Please notice the logarithmic scale on the ordinate for analogical density (nano (n): 10^{-9}, pico (p): 10^{-12}, femto (f): 10^{-15}.). The tokenisation schemes on the abscissae are sorted according to the average length of tokens in ascending order.

ber of analogies (top-right). Tatoeba has about 10 times more analogies than Multi30K.

Figure 2 (bottom-left) shows the results on the analogical density of the corpora with various tokenisations. We can immediately observe that Tatoeba corpus steadily has the highest analogical density in comparison to the other corpora. The difference is also pretty far. For example, the gap is around 10^3 between Tatoeba and Multi30K, even more than 10^5 for Europarl. This shows that Tatoeba corpus is really denser than the other corpora despite having the smallest number of sentences. Remember, we have different numbers of sentences between corpora.

Although it is not visible from the graph, we observed that the density slightly gets lower from tokenisation in character towards words. For subword tokenisation, we found that unigram consistently has higher analogical density than BPE on the same vocabulary size. This is probably caused by the unigram having a shorter token length which allows a higher degree of freedom in commutation between tokens.

Similar trends can also be observed in the proportion of analogical sentences. Tatoeba is ten times higher than Multi30K which proves our hypothesis that a corpus which contains similar sentences will have a higher proportion of analogical sentences. As for the tokenisation scheme, we also found that the proportion decreases toward word tokenisation.

5 Conclusion

We produced a resource of analogies between sentences extracted from two different corpora, Tatoeba and Multi30K. Both corpora have different characteristics, the one contains mostly daily life conversations, and the other contains a collection of image captions. We also performed experiments in measuring the analogical density of various corpora in various languages using different tokenisation schemes. Corpora with a higher Type-Token-Ratio tend to have higher analogical densities. We naturally found that the analogical density goes down from character to word. We hope the release of such a resource will allow a better way to evaluate the quality of sentence embeddings.

Acknowledgment

This work was supported by a JSPS grant, number 18K11447 (Kakenhi Kiban C), entitled ”Self-explainable and fast-to-train example-based machine translation”.

References

[1] Yves Lepage. Solving analogies on words: an algorithm. In *Proceedings of the 17th international conference on Computational linguistics (COLING 1998)*, Vol. 1, pp. 728–734. Association for Computational Linguistics, 1998.

[2] Nicolas Stroppa and François Yvon. An analogical learner for morphological analysis. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pp. 120–127, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[3] Philippe Langlais and François Yvon. Scaling up analogical learning. In *Coling 2008: Companion volume: Posters*, pp. 51–54, Manchester, UK, August 2008. Coling 2008 Organizing Committee.

[4] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pp. 70–74. Association for Computational Linguistics, 2016.

[5] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pp. 215–233, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[6] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 304–323, 2018.

[7] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, Vol. 2, pp. 67–78, 2014.

[8] Rashel Fam and Yves Lepage. Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC-18)*, pp. 1060–1066, Miyazaki, Japan, May 2018. ELRA.

[9] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[10] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.