A Multi-Domain Framework for Textual Similarity.
A Case Study on Question-to-Question and Question-Answering Similarity Tasks

Amir Hazem, Basma El Amal Boussaha and Nicolas Hernandez
Laboratoire des Sciences du Numérique de Nantes (LS2N)
Université de Nantes (France)
Amir.Hazem@univ-nantes.fr

LREC 2018 MIYAZAKI

Context

- Community Question Answering (CQA)
- Task1→question-to-question similarity (Q/Q)
- Task2→question-answering similarity (Q/A)
- The first multi-domain framework for Q/Q and Q/A tasks StackExchange
- Evaluation of 2 baselines over 19 domains

Datasets

- A multi-domain CQA framework
- 19 StackExchange datasets
- Based on a voting system and the tag "**Answer**" provided by the metadata
- Comments are not considered

Corpus Size

Corpus	#token	#all posts	#filtered posts	#test
Earth Science	221K	2.2k	1.6k	169
Expatriates	185k	2.6k	1.3k	137
Health	276k	2.9k	2.2k	223
Sports	264K	3.2k	2.3k	240
Politics	415K	3.2k	2.8k	282
Pets	373k	3.4k	2.5k	253
Economics	333k	4.1k	2.1k	210
Law	609k	5.1k	3.6k	365
History	741k	6.1k	5.2k	522
Philosophy	1.1M	7.3k	5.7k	575
Music	701k	9.1k	5.4k	544
Workplace	1.7M	12.9k	8.2k	830
Biology	1.1M	14.1k	10k	1001
Cooking	1.2M	16k	11.3k	1132
Chemistry	1.3M	18.5k	10.4k	1042
Travel	1.6M	20.6k	12.9k	1297
Physics	7.02M	87.2k	44.4k	4443
AskUbuntu	11.1M	248k	79.1k	7912
Math	28.8M	702k	168k	16820

Table 1: Size of the multi-domain datasets in terms of number of tokens (column 1), number of posts (column 2), number of filtered posts (column 3) and number of test questions (column 4).

Approaches

SentEmb

- Q/A Vectors: sum of the embeddings of each word of Q/Q and Q/A
- Cosine to extract similar pairs of Q/Q or Q/A
- Pre-processing (TOK + POSTAG + LEM)
- Stop-words are removed, only nouns, verbs and adjectives are kept

MappSent

- Extension of SentEmb
- Build new Q/Q and Q/A embedding spaces
- SVD to build a mapping matrix
- Project pairs of Q/Q, Q/A in a new subspace
- Cosine to extract similar pairs of Q/Q or Q/A

Q/A Example

Original question (Subject):
What 's an appropriate diet for a small river turtle?

Related question (Body):
I get my little brother a small river turtle, and i would like to know what it eat. it's somewhat relatively fast, longish slim tail, seem to be energetic, but i 'm not sure what it eat. Any idea?

Answer:
In general a turtle's diet is comprised of 50% protein, 30% vegetable, 10% green, and 10% fruit. The protein can even be divided up so that it's 25% commercial food, and 25% fresh food. You should feed the turtle once daily since it's so young right now, but as it get old you can move to a schedule of every other day. Some people even go so low every third day; others keep feed them daily but in small amount. It's important to keep track of how much you're feeding your turtle because turtle is notorious beggar, and you can risk get your turtle fat if you give in to it demand too often. The rule of thumb for feed time is as much as the turtle can eat in a 15-minute time-span, or an amount of food that match the size of the turtle's head and neck.

Results

Corpus	*SentEmb*		*MappSent*	
	Dev	Test	Dev	Test
Earth Science	**67.7**	68.2	67.4	**73.1**
Expatriates	65.7	68.2	**68.8**	**71.1**
Health	45.8	66.3	**46.5**	**66.8**
Sports	62.7	57.5	**64.3**	**59.7**
Politics	70.4	70.3	**73.3**	**72.0**
Pets	60.9	61.6	**63.7**	**63.2**
Economics	66.2	61.0	**67.2**	**61.4**
Law	60.7	**71.1**	**62.1**	70.8
History	51.7	60.3	**52.9**	**62.7**
Philosophy	35.9	40.8	**40.3**	**44.6**
Music	46.9	44.1	**49.2**	**45.9**
Workplace	40.4	39.6	**43.1**	**41.5**
Biology	50.0	37.4	**52.8**	**38.9**
Cooking	54.0	50.7	**57.1**	**53.2**
Chemistry	37.0	41.2	**38.9**	**43.6**
Travel	53.9	53.8	**56.6**	**57.2**
Physics	37.1	32.4	**40.1**	**34.5**
AskUbuntu	13.6	18.5	**14.7**	**19.8**
Math	6.23	5.83	**6.71**	**6.13**

Table 2: Results (MAP%) of SentEmb and MappSent on the question-to-question similarity task using 19 Q/Q datasets.

Corpus	*SentEmb*		*MappSent*	
	Dev	Test	Dev	Test
Earth Science	16.9	9.01	**41.4**	**41.2**
Expatriates	7.02	5.35	**26.0**	**25.9**
Health	4.63	4.25	**24.4**	**22.4**
Sports	10.0	11.4	**42.2**	**33.5**
Politics	8.09	6.60	**32.4**	**36.1**
Pets	9.64	9.66	**27.3**	**33.2**
Economics	9.29	4.44	**32.5**	**27.7**
Law	6.24	5.89	**26.7**	**25.2**
History	7.45	8.47	**33.0**	**33.4**
Philosophy	6.03	5.44	**26.1**	**22.1**
Music	12.1	11.4	**25.7**	**27.7**
Workplace	9.71	**16.1**	**14.3**	13.4
Biology	2.32	1.85	**28.1**	**27.8**
Cooking	9.66	3.31	**25.9**	**27.1**
Chemistry	2.01	2.97	**17.0**	**18.1**
Travel	3.75	5.37	**23.4**	**24.8**
Physics	1.07	1.17	**13.7**	**14.1**
AskUbuntu	0.45	0.29	**4.88**	**5.58**
Math	0.21	0.12	**4.07**	**6.43**

Table 3: Results (MAP%) of SentEmb and MappSent on the Question-Answering similarity task using 19 Q/A datasets.

Conclusion

- We provide more realistic multi-topic datasets to evaluate systems performance on textual similarity tasks
- We targeted question-to-question and question-answering similarity tasks which represent a key challenge in community question answering
- We provide the first version of 19 raw and pre-processed datasets of various topics
- These datasets will be gradually extended and enriched in the near future to provide the 180 datasets contained in StackExchange
- We hope that this work will serve as a basis for future evaluation on these tasks

Acknowledgments

The current work was supported by the ANR 2016 PASTEL (ANR-16-CE33-0007) project