Making Test Corpora for Question Answering More Representative

Andrew Walker[1], Andrew Starkey[2], Jeff Z. Pan[1], and Advaith Siddharthan[1]

[1] Computing Science, University of Aberdeen, UK
{r05aw0,jeff.z.pan,advaith}@abdn.ac.uk
[2] Engineering, University of Aberdeen, UK
a.starkey@abdn.ac.uk

Abstract. Despite two high profile series of challenges devoted to question answering technologies there remains no formal study into the representativeness that question corpora bear to real end-user inputs. We examine the corpora used presently and historically in the TREC and QALD challenges in juxtaposition with two more from natural sources and identify a degree of disjointedness between the two. We analyse these differences in depth before discussing a candidate approach to question corpora generation and provide a juxtaposition on its own representativeness. We conclude that these artificial corpora have good overall coverage of grammatical structures but the distribution is skewed, meaning performance measures may be inaccurate.

1 Introduction

Question Answering (QA) technologies were envisioned early on in the artificial intelligence community. At least 15 experimental English language QA systems were described by [13]. Notable early attempts include BASEBALL [11] and LUNAR [17,18]. New technologies and resources often prompt a new wave of QA solutions using them. For example: relational databases [8] with PLANES [16]; the semantic web [2] by [3]; and Wikipedia [15] by [7].

Attempts to evaluate QA technologies are similarly diverse. The long-running Text REtrieval Conferences[1] (TREC) making use of human assessors in conjunction with a nugget pyramid method [12], while the newer Question Answering over Linked Data[2] (QALD) series uses an automated process that compares results with a gold standard.

In both cases, however, the matter of whether or not the questions being posed to the challenge participants actually capture the range and diversity of questions that real users would make of a QA system is not addressed. We explore the distribution of grammatical relationships present in various artificial and natural question corpora in two primary aspects: coverage and representativeness. Coverage is important for QA solution developers to gauge gaps in their

[1] http://trec.nist.gov/

[2] http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/

E. Kanoulas et al. (Eds.): CLEF 2014, LNCS 8685, pp. 1–6, 2014.

system's capacity, whereas evaluations are dependent on the representativeness of their corpora for valid comparisons between systems.

2 Corpora Sources

To evaluate a QA system for commercial use, it would be preferable to test it on real user questions. That is, questions that have been posed by potential or real end-users rather than system developers or testers. Although artificial questions may be used to capture additional grammatical forms, the most important aspect for a functional QA system is to answer those put by real users.

We collected questions from 4 distinct corpora, 2 artificial and 2 natural:

1. TREC has been running since 1992 and published 2,524 unique questions with which to evaluate text retrieval system submissions. These questions are artificial by the track organisers and often pertain to given contexts not found in the questions themselves. For example, a question "What was her name?" makes sense within a context, but is essentially meaningless alone.
2. The QALD challenges have been running since 2011, publishing 453 unique questions focussed on DBpedia [1,4] and MusicBrainz [14] data. These also are artificial but are always context independent.
3. We extracted 329,510 questions from Yahoo! Answers[3] tagged as English. These are the question titles put by the general public for other members of the public to propose answers to, and so in some cases do not form typical question structures – leaving the details of the question to the post's body.
4. A set of 78 questions put by participants of OWL tutorials to a Pizza ontology. These are considered natural as the participants were not experts and the questions include some grammatical and spelling errors.

3 Analysis and Comparison of Question Corpora

We seek to compare the entries of the various corpora in order to discern if the artificial questions currently being used for QA system evaluation are representative of the questions real end-users might pose. If some feature or aspect of natural language questions are over- or under-represented in an evaluation corpus this will cause evaluation measurements to be inaccurate as accounts of a QA system's performance in an end-system.

Rather than manually inspecting the grammatical forms of all 332,565 entries, we ran a statistical analysis comparing the distribution of various grammatical relations found in the corpora. Using the Stanford Parser[4] [10,9] we derived the dependency graph for each question and then, for each corpus D, computed frequency vectors for each dependency type t, normalised by *tf-idf*[5]. We then

[3] http://answers.yahoo.com/

[4] Stanford CoreNLP version 1.3.5 trained with the provided English PCFG model

[5] Term frequency - inverse document frequency
$\mathrm{tfidf}\left(t,D\right)=\log\left(\mathrm{f}\left(t,D\right)+1\right)\times\log\frac{N}{\left|\left\{d\in D:t\in d\right\}\right|}$ where t is a dependency type and D is a corpus of questions.

compared the distribution of dependency types across the four corpora in two ways: by calculating pairwise cosine similarity, and by calculating pairwise Pearson correlation between corpora. These comparisons are shown in Table 1.

Table 1. Pairwise comparison of dependency type distributions across corpora

Yahoo!	**Pizza**	**QALD**	
0.7847	0.7479	0.8593	**TREC**
	0.6060	0.7428	**Yahoo!**
		0.8373	**Pizza**

(a) Cosine similarity

Yahoo!	**Pizza**	**QALD**	
−0.1018	0.2350	**0.4942**	**TREC**
	-0.5492	−0.2023	**Yahoo!**
		0.5345	**Pizza**

(b) Pearson. **Bold** indicates $p < 0.05$

Of note in Table 1a is the strong similarity of distributions between the two artificial corpora, QALD and TREC, where comparisons with them and the natural corpora show weaker correspondence. The Yahoo! corpus shows relatively low similarity with any other corpus – perhaps due to its heavy reliance on colloquialisms and overwhelming prominence of ungrammatical content. Table 1b emphasises the dissimilarity of Yahoo! to the other sources.

Table 2. Dependency relations that are over- and under-represented in artificial corpora. Discussed relations are in **bold**. A * indicates possible over-representation, and a † under-representation.

relation	**trec**	**pizza**	**qald**	**yahoo**
det*	17.63	13.66	18.19	9.36
prep*	12.52	9.95	12.90	10.30
nn*	10.56	3.47	12.09	7.28
aux†	4.38	4.86	2.68	7.57
dep†	2.08	6.48	2.72	5.73
attr*	5.52	0.69	2.68	1.36
conj†	0.64	5.09	0.51	2.67
cop†	0.63	3.24	1.18	2.15
auxpass*	3.16	0.69	2.54	0.63
nsubjpass*	3.16	0.69	2.57	0.55
xcomp†	0.60	1.62	0.44	1.99
appos†	0.26	0.46	0.04	0.39
neg†	0.02	0.46	0.07	0.58
agent*	0.37	0	0.59	0.07
ccomp†	0.20	1.16	0.22	2.13
mark†	0.14	0.46	0.11	1.26
advcl†	0.15	0.23	0.04	1.08
csubj†	0.01	0.23	0.18	0.24
predet†	0.01	0.46	0	0.08
cc†	0	0.23	0.07	0.13
preconj†	0	0.23	0	0.01

There are some grammatical dependency relations that are interesting in their under-representation within artificial question corpora.

The `predet` relation (predeterminer) is found only twice in TREC and never in QALD, but enjoys greater usage in the Pizza and Yahoo! corpora. This relation is typically found connecting a noun and the word “all”, as in “Find all the pizzas with less than 3 toppings”.

Similarly, the preconj relation (preconjunct) is never found in QALD or TREC, but has limited exposure in both Pizza and Yahoo! corpora. This is a relation "between the head of [a noun-phrase] and a word that appears at the beginning bracketing a conjunction (and puts emphasis on it), such as 'either', 'both', 'neither')", as in "Which pizza has neither vegetables nor seafood?".

The artificial corpora contain many more attr relations, which are used to attach the main verb to a leading *wh-* word, suggesting that the corpora authors are relying too heavily on *wh-* formulations.

The nn relation (noun compound modifier) sees heavy use in both TREC and QALD but is not similarly represented in the natural Pizza and Yahoo! corpora. This may be due in part to domain dependence, with questions focussed on named entities.

4 Constructing Evaluation Corpora

Having established that artificial and natural corpora of natural language questions have discrepancies in grammatical form and variation, we ask how one might compose an evaluation corpus of such questions for a given domain that maintains representativeness of real end-user inputs. We suggest a lexical substitution approach, taking examples from natural question sets and replacing mappable concepts with those from the required domain. This is applied to two scenarios: first, with a case study of QALD seeking to improve its representativeness in Sect. 4.1 and second on building a new corpus from scratch in Sect. 4.2. It is the corpus engineer's responsibility to ensure sensible substitutions.

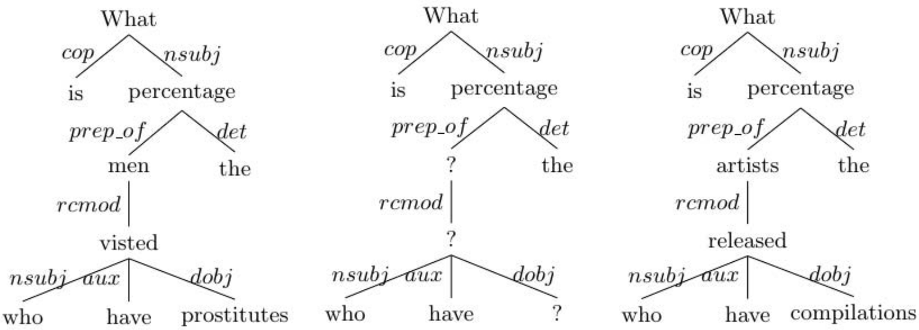
4.1 Extending QALD to Improve Representativeness

For this section we will be using the QALD-3 DBpedia testing corpus, which consists of 100 questions collectively bearing a 0.5952 cosine similarity with the Yahoo! Answers corpus, in terms of *tf-idf* distribution.

We draw entries from Yahoo! at random and calculate the effect its inclusion would have on the cosine similarity score. When a positive effect is found, that entry is examined for suitability. For any with suitable dependency graphs, we apply lexical substitution to render the question appropriate for the target ontology while maintaining the original grammatical structure.

For example, imagine that the question "what is the percentage of men who have visted$^{[sic]}$ prostitutes?"[6] was one selected in this manner. We can identify the core concepts of the question and substitute them with concepts and instances from MusicBrainz. In this case we could choose "What is the percentage of artists who have released compilations?", as shown in Fig. 1. Figure 2 shows the growth of similarity score with just a few iterations of this method. This process can be repeated as desired to build a corpus of relevant questions with more representative distributions of grammatical dependencies.

[6] Although this entry contains typographic errors, the parser nevertheless gives a usable dependency graph.



(a) original (b) anonymised (c) substituted

Fig. 1. Lexical substitution within a question by dependency graph [7]

4.2 Building a New Evaluation Corpus

The strategy also applies to the construction of entirely new corpora. We would initially choose questions that individually bear the greatest similarity to Yahoo! as a whole and then reiterate with the process as before, for extending an existing corpus.

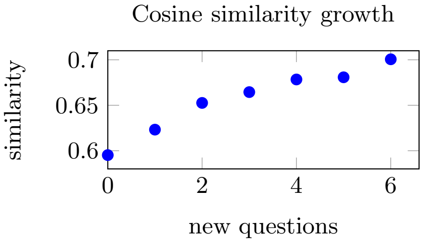


Fig. 2. Similarity growth with new questions

5 Conclusions and Future Work

Two corpora of artificial English questions demonstrate stronger similarity with each other than either of two corpora of natural English questions. This suggests that results of evaluations of QA systems using these artificial corpora may not be indicative of performance on natural questions. We proposed a methodology for creating natural questions within a domain by performing lexical substitution within samples of natural-provenance questions from other domains.

This study pertains to low-level analysis of English questions and does not address coverage and representativeness of other linguistic features. Although substitutions are tailored to a given context, no effort is made explicit here to emulate the distribution of question topics; this should be the responsibility of the corpus engineer.

[7] For conciseness we use the collapsed graphs using "`prep_of`" but this has no bearing on the result.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)
2. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific American 284(5), 28–37 (2001)
3. Bernstein, A., Kaufmann, E., Göhring, A., Kiefer, C.: Querying ontologies: A controlled english interface for end-users. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 112–126. Springer, Heidelberg (2005)
4. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia-a crystallization point for the web of data. Web Semantics: Science, Services and Agents on the World Wide Web 7(3), 154–165 (2009)
5. Brill, E., Lin, J., Banko, M., Dumais, S., Ng, A., et al.: Data-intensive question answering. In: Proceedings of the Tenth Text REtrieval Conference, TREC 2001 (2001)
6. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems 30(1), 107–117 (1998)
7. Buscaldi, D., Rosso, P.: Mining knowledge from wikipedia for the question answering task. In: Proceedings of the International Conference on Language Resources and Evaluation (2006)
8. Codd, E.F.: A relational model of data for large shared data banks. Communications of the ACM 13(6), 377–387 (1970)
9. De Marneffe, M.C.: What's that supposed to mean? Ph.D. thesis, Stanford University (2012)
10. De Marneffe, M.C., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: Proceedings of LREC, vol. 6, pp. 449–454 (2006)
11. Green, Jr., B.F., Wolf, A.K., Chomsky, C., Laughery, K.: Baseball: an automatic question-answerer. Papers Presented at the May 9-11, 1961, western joint IRE-AIEE-ACM Computer Conference, pp. 219–224. ACM (1961)
12. Lin, J., Demner-Fushman, D.: Will pyramids built of nuggets topple over? In: Proceedings of the main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp. 383–390. Association for Computational Linguistics (2006)
13. Simmons, R.F.: Answering english questions by computer: a survey. Commun. ACM 8(1), 53–70 (1965), http://doi.acm.org/10.1145/363707.363732
14. Swartz, A.: Musicbrainz: A semantic web service. IEEE Intelligent Systems 17(1), 76–77 (2002)
15. Wales, J., Sanger, L.: Wikipedia, the free encyclopedia (2001), http://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=551616049 (accessed April 22, 2013)
16. Waltz, D.L.: An english language question answering system for a large relational database. Communications of the ACM 21(7), 526–539 (1978)
17. Woods, W.A.: Progress in natural language understanding: an application to lunar geology. In: Proceedings of the National Computer Conference and Exposition, AFIPS 1973, June 4-8, 1973, pp. 441–450. ACM, New York (1973), http://doi.acm.org/10.1145/1499586.1499695
18. Woods, W.A.: Lunar rocks in natural english: Explorations in natural language question answering. Linguistic Structures Processing 5, 521–569 (1977)