With a little help from NLP: My Language Technology applications with impact on society

Ruslan Mitkov

University of Wolverhampton

Abstract

The keynote speech presents the speaker's vision that research should lead to the development of applications which benefit society. To support this, the speaker will present three original methodologies proposed by him which underpin applications jointly implemented with colleagues from across his research group. These Language Technology tools already have a substantial societal impact in the following areas: learning and assessment, translation and care for people with language disabilities.

1. Impact on learning and assessment

The first part of the presentation will introduce an original methodology and tool for generating multiple-choice tests from electronic documents. Multiple-choice tests are sets of test items, the latter consisting of a question or *stem* (e.g. Who was FIFA player of the year for 2017?), the correct *answer* (e.g. Ronaldo) and *distractors* (e.g. Messi, Neymar, Buffon). This type of test has proved to be an efficient tool for measuring students' achievement and is used on a daily basis both for assessment and diagnostics worldwide. According to Question Mark Computing Ltd (p.c.), who have licensed their Perception software to approximately three million users so far, 95% of their users employ this software to administer multiple-choice tests. Despite their popularity, the manual construction of such tests remains a time-consuming and labour-intensive task. One of the main challenges in constructing a multiple-choice test item is the selection of plausible alternatives to the correct answer which will better distinguish confident students from unconfident ones.

As an illustration, consider the sentence "Syntax is the branch of linguistics which studies the way words are put together into sentences". This sentence can be transformed into the questions "Which branch of linguistics studies the way words are put together into sentences?", "Which discipline studies the way words are put together into sentences?" or "What studies the way words are put together into sentences?". All these phrases can act as stems in multiple-choice test items. If we assume that the stem of a test item is one of the questions above, the distractors to the correct answer *syntax* should preferably be concepts semantically close to it. This is vital because in this case the distractors will be more plausible and therefore better at distinguishing good, confident students from poor and uncertain ones. For this particular test item, *semantics* or *pragmatics* would be a much better distractors than *chemistry* or *football*, for instance.

Mitkov and Ha (2003) and Mitkov et al. (2006) offered an alternative to the lengthy and demanding activity of developing multiple-choice test items by proposing an NLP-based methodology for construction of test items from instructive texts such as textbook chapters and encyclopaedical entries. This methodology makes use of NLP techniques including shallow parsing, term extraction, sentence transformation and semantic distance computing and employs resources such as corpora and ontologies like WordNet. More specifically, the system identifies important terms in a textbook text, transforms declarative sentences into questions and mines for terms which are semantically close to the correct answer, to serve as distractors.

The system for generation of multiple-choice tests described in Mitkov and Ha (2003) and in Mitkov et al. (2006) was evaluated in practical environment where the user was offered the option to post-edit and in general to accept or reject the test items generated by the system. The formal

Keywords: NLP-based methods, learning and assessment, translation, language disabilities

evaluation showed that even though a significant part of the generated test items had to be discarded, and that the majority of the items classed as 'usable' had to be revised and improved by humans, the quality of the items generated and proposed by the system was not inferior to the tests authored by humans, were more diverse in terms of topics and very importantly – their production needed 4 times less time than the manually written items. The evaluation was conducted both in terms of measuring the time needed to develop test items and in terms of classical test analysis to assess the quality of test items.

A later study (Mitkov et al. 2009) sought to establish which similarity measures generate better quality distractors of multiple-choice tests. Similarity measures employed in the procedure of selection of distractors were collocation patterns, four different methods of WordNet-based semantic similarity (extended gloss overlap measure, Leacock and Chodorow's, Jiang and Contrath's as well as Lin's measures), distributional similarity, phonetic similarity as well as a mixed strategy combining the aforementioned measures. The evaluation results showed that the methods based on Lin's measure and on the mixed strategy outperform the rest, albeit not in a statistically significant fashion.

The system for generation of multiple-choice tests has been taken up by the National Board of Medical Examiners (NBME) based in Philadelphia, USA. NBME are the only organisation in USA who are licenced to administer and asses exams for the to-be-doctors. NBME have been been using our system for delivery of low-stake tests for more than 10 years already.

2. Impact on translation

The quest for reliable tools assisting professional translators goes back to 1971 when Krollman (1971) put forward the reuse of existing human translations. A few years later, Arthern (1979) went further and proposed the retrieval and reuse not only of identical text fragments (exact matches) but also of similar source sentences and their translations (fuzzy matches). It took another decade before the ideas sketched by Krollman and Arthern were commercialised as a result of the development of various computer-aided translation (CAT) tools such as Translation Memory (TM) systems in the early 1990s. These translation tools revolutionised the work of translators and the last two decades saw dramatic changes in the translation workflow.

The TM memory systems indeed revolutionised the work of translators and nowadays the translators not benefiting from these tools are a tiny minority. However, while these tools have proven to be very efficient for repetitive and voluminous texts, are they intelligent enough? Unfortunately, they operate on fuzzy (surface) matching (Levenstein distance) mostly, cannot reuse already translated sentences which are part of another complex sentence nor texts which are synonymous to (or paraphrased versions of) the text to be translated and can be 'fooled' on numerous occasions. A recent study (Mitkov et al., forthcoming) shows that TM systems (Trados, MemoQ, Wordfast, Omega T) spectacularly fail to offer matches for sentences already translated but which have undergone (slight) transformations which include among others: change active to passive voice and vice versa, change word order and replace one word with synonym.

A way forward would be to equip the TM tools with Natural Language Processing (NLP) capabilities. This idea was suggested first by Mitkov (2005) at panel discussion held during 27th annual conference Translating and the Computer in London and the first experiments were reported by Mitkov and Pekar (2007). In the second part of his presentation, the speaker will explain how two NLP methods/tasks, namely clause splitting and paraphrasing, make it possible for TM systems to identify semantically equivalent sentences which are not necessarily identical or close syntactically and enhance performance. The results reported in Timonera and Mitkov (2015) show that TM systems which are enhanced with a clause splitting component perform with a dramatic increase of recall which is statistically significant. Adding a paraphrasing module increases further the performance and experiments with the S to XL package sizes of the Paraphrase Database PPDB show that the larger the database, the better the results.

In (Gupta et al. 2016) we presented a novel and efficient approach which incorporates semantic information in the form of paraphrasing in the edit-distance metric. The approach computes edit-distance while efficiently considering paraphrases using dynamic programming and greedy approximation. In addition to using automatic evaluation metrics such as BLEU and METEOR, we have carried out an extensive human evaluation in which we measured post-editing time, keystrokes, HTER, HMETEOR, and carried out three rounds of subjective evaluations. Our results show that

paraphrasing substantially improves TM matching and retrieval, resulting in translation performance increases when translators use paraphrase-enhanced TMs. Finally, the speaker will present a new metric developed by members of his group (Gupta et al. 2014) which is capable of comparing semantic similarity of sentences and thus becomes highly eligible for inclusion in a new generation TM matching algorithm.

The speaker will promise to go beyond the translation world: he is already thinking not only about the next-generation translation memory tools for translators but also about the future interpreting memory tools for interpreters. The presentation will sketch how this is envisaged to be developed and how it will work. In addition to the interpreting memory, the speaker will outline other tools which will be developed as support to interpreters.

3. **Impact on people with language disabilities**

The last part of the keynote speech will focus on the work within the recent EC-funded project FIRST whose objective was to develop a tool customised for the needs of people with autism (ASD) by allowing easy comprehension of texts which otherwise would have been challenge for them (Mitkov 2011; Orasan et al. 2012; Orasan et al. 2017). Autistic Spectrum Disorder (ASD) is a neurodevelopmental disorder which has a life-long impact on the lives of people diagnosed with the condition. In many cases, people with ASD are unable to derive the gist or meaning of written documents due to their inability to process complex sentences, understand non-literal text, and understand uncommon and technical terms. The idea put forward by the speaker was to develop a tool which would enable readers or carers to convert documents into easier-to-understand ones by (i) reducing complexity at morphological and syntactical level, (ii) by removing ambiguity in terms of lexical polysemy, anaphoric interpretation and figurative language and (iii) by improving readability through adding pictures, document navigation tools, providing concise summaries of long documents and replacing technical words with more common ones. The project FIRST produced a powerful editor called OpenBook which is operational for English, Spanish and Bulgarian and which enables carers of people with ASD to prepare texts suitable for this population. Assessment of the texts generated using the editor showed that they are not less readable than those generated more slowly as a result of onerous unaided conversion and were significantly more readable than the originals.

The speaker intends to go beyond the topic of autism and plans to develop tools customised for people with dementia. The first goal of this project will consist of the data collection of speech samples to build a corpus of transcribed speech of Alzheimer's disease and control subjects. Language technology and machine learning techniques will be employed to measure a set of speech and language markers and to assess their change. The project is expected to contribute to the understanding of changes in language use of people with dementia. It will also enhance understanding of communication in this population and will suggest improved therapeutic strategies involving the use of language and information technology to automatically correct some of the communication deficiencies identified in people with dementia.

References

Arthern, P. J. 1979. "Machine Translation and computerized terminology systems: A translator's viewpoint". In *Translating and the computer, proceedings of a seminar. London, 14th November, 1978*, ed. Snell, B. M., 77-108. Amsterdam: North-Holland.

Gupta, R., Bechara, H., El Maarouf, I. and Orasan, C. 2014. "UoW: NLP techniques developed at the University of Wolverhampton for Semantic Similarity and Textual Entailment". In *Proceedings of the 8th International Workshop on Semantic Evaluation* (SemEval 2014), 785-789. Dublin, Ireland.

Gupta, R., Orasan, C., Zampieri, M., Vela. M., Mihaela Vela, van Genabith, J. and Mitkov, R. 2016. "Improving Translation Memory matching and retrieval using paraphrases", *Machine Translation*, 30(1), 19-4

Gupta, R., Orasan, C., Liu, Q. and Mitkov, R. 2016. A Dynamic Programming Approach to Improving Translation Memory Matching and Retrieval using Paraphrases. In *Proceedings of the 19th International Conference on Text, Speech and Dialogue (TSD)*, Brno, Czech Republic.

Krollmann, F. 1971. "Linguistic data banks and the technical translator". *Meta* 16(1-2), 117-124.

Mitkov, R. 2011. A Flexible Interactive Reading Support Tool (FIRST). *Presentation at the FIRST Kick-Off Meeting*. Brussels, 17th October 2011.

Mitkov, R. 2005. "New Generation Translation Memory systems". *Panel discussion at the 27th annual Aslib conference 'Translating and the Computer'*. London.

Mitkov, R. and Ha, L.A. 2003. "Computer-aided generation of multiple-choice tests". *Proceedings of the HLT/NAACL 2003 Workshop on Building educational applications using Natural Language Processing*, 17-22. Edmonton, Canada.

Mitkov, R., An, L.A. and Karamanis, N. 2006. "A computer-aided environment for generating multiple-choice test items". *Journal of Natural Language Engineering*, 12 (2), 177-194.

Mitkov, R., Ha, L.A., Varga, A. and L. Rello. 2009. "Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation". *Proceedings of the EACL'2009 workshop on geometric models of natural language semantics*, 49-56. Athens, Greece

Mitkov, R., Feherova A., Cotella, M., and Silvestre Baquero, A. (forthcoming) "Translation Memory systems have a long way to go"

Orasan, C., Evans, R., and Dornescu, I. (2013) Text Simplification for People with Autistic Spectrum Disorders. In Tufis, D., Rus, V. and Forascu, C. (eds.), Towards Multilingual Europe 2020: A Romanian Perspective, Romanian Academy Publishing House, Bucharest, 2013, ISBN: 978-973-27-2282-4, pp. 287-312.

Orasan, C., Evans, R. and Ruslan Mitkov, R. 2018. "Intelligent Text Processing to Help Readers with Autism". *Intelligent Natural Language Processing: Trends and Applications*. Springer.

Pekar V. and Mitkov R. 2007. "New Generation Translation Memory: Content-Sensitive Matching". *Proceedings of the 40th Anniversary Congress of the Swiss Association of Translators, Terminologists and Interpreters*. Bern.

Timonera, K. and R. Mitkov. 2015. Improving Translation Memory Matching through Clause Splitting. *Proceedings of the RANLP'2015 workshop 'Natural Language Processing for Translation Memories'*. Hissar, Bulgaria.