Machine Evaluation of Analytic Products
Detailed Description

I. Introduction

Modern machine learning techniques can sometimes provide sufficient performance without taking in to account any domain specific concerns, but it must be noted that there is no one-size-fits-all. In the process of implementing what follows, it may well be that the method intended for *solution requirement 3* will end up working better for *solution requirement 6*, while *solution requirement 8* could see the best performance when combining the tactics used in several other sections, and so on. Regardless, all the pieces necessary for a prototype are present, and the quality of evaluations is expected to approach the same level as a very distracted expert – the software system described herein is far beyond human speeds, but it will not be capable of matching a human in other ways. Without further details regarding expected use cases and necessary minimum requirements (i.e. how skilled current humans are, in practice, at performing this task), I'm not able to say that the system proposed here will be any better than sufficient for all expected purposes.

II. Solution Requirements

1. Properly describes quality and credibility of underlying sources, data, and methodologies

The fundamentals of this task can be best handled without any methods that require supervised training (manually labeled data.) Part of speech tagging and named entity extraction can generate a list of expected references and/or sources, and compare the *implicit* citations of the analytic product against the expected list. *Explicit* citations are, of course, even easier.

A recent method that preserves the unsupervised nature of part of speech tagging (data need not be manually labeled), while also providing much of the power of supervised methods, is cascading part of speech tagging (can also be called keyphrase extraction.) In essence, this method maintains a minimal amount of metadata with which to describe and compare named entities; as text is processed, the metadata is used to compare linguistic sequences. Much more complex linguistic constructs can be accurately recognized this way. [3]

2. Properly expresses and explains uncertainties associated with major analytic judgments

Sentiment analysis is used here to compare the explicit statements contained within the analytic product, against their sentiment. The reader should note that sentiment analysis is not restricted to dealing with the perceived tone of a sample; it can also be used as an indicator of the desirability of semantic content. To accomplish this, ontological features are also checked and considered. [5]

3. Properly distinguishes between underlying intelligence information and analysts' assumptions and judgments

The sentiment analysis and ontological features previously mentioned are used here as well; additionally, an abstract meaning representation design, which incorporates attention mechanisms via a pointer network (implemented using a long short term memory model), is required. While much larger, slower, and narrower in scope implementations can be applied once the application domain is held fixed, the pointer network method is light weight by comparison and much broader in scope. [2] [5]

4. Incorporates analysis of alternatives

To accomplish this task, all previously mentioned designs are required. Also of use is an ensemble method combining naive bayes, decision tree, and support vector machine designs; this ensemble framework was originally conceived of in order to evaluate rumors found throughout social media content. It should, with a different training mechanism, prove quite effective in addressing the holes that would otherwise prevent this solution requirement from being satisfied. [6] [5] [2]

5. Demonstrates relevance to customers and addresses implications and opportunities

I can think of no way to address this solution requirement without further information; a number of possibilities exist, but none of them are omniscient. Without knowing the scope of the demands of customers, I also cannot state with any certainty that a satisfactory method – that is, one which considers the full range of relevance – is even publicly known at this time.

6. Uses clear and logical argumentation

Comparing and contrasting event representations, in addition to the abstract meaning representation methods already discussed, will satisfy this requirement. Note that thorough results for DARPA's DEFT project are not currently available; thus, none of the representations found in the cited works may be satisfactory. A proof-of-concept would be required for this to be stated definitively. [1]

7. Explains change to or consistency of analytic judgments

A gold-standard approach, i.e. maintaining a corpus of analytic products that are representative of the gold-standard within their specific area(s) of coverage, will provide the best performance here. The corpus need not be fixed, and a constantly changing database would work quite nicely. Document graphs and query based summarization are ideal designs for this purpose. [4]

8. Makes sound judgments and assessments

A variation of the gold-standard approach previously mentioned will satisfy this requirement; [4] what is even more likely to be of use, however, is a variation of one (or more) of the methods set forth in the Xpress challenge. The details and submissions for the Xpress challenge are not known by this author.

9. Incorporates effective visual information where appropriate

A method which will work in many cases is a fully convolutional network with attention mechanisms, trained to allow for posing short questions. [7]

III. Prototype development

A. Timeframe

It would take a competent team 2-6 months to have a working prototype finished; a single expert could likely do all of the work in about six months. Note that *prototype* is the operative word here; a finished product would take longer.

B. Budget

The only expenses would be in the form of overhead costs and engineer compensation (which is likely to be higher than this author expects.) Hardware costs would be almost nonexistent, since engineers able to do the job would almost certainly have workstations of their own (i.e. high end PCs.) If proprietary code is desired, however, then it would need to be purchased or implemented from scratch. That could cost up to several million dollars, or take up to two years of additional time. Open source software is really the way to go here.

C. Other necessary resources

A very large amount of analytic products would be necessary. To avoid security concerns, the vast majority could be out-of-date products, but at least 500 examples would be necessary (more would be better.) They should, at a minimum, be a representative sample of the variations that are seen in practice (subject, intended audience, etc.)

References

[1] Ann Bies, Zhiyi Song, Jeremy Getman, Joe Ellis, Justin Mott, Stephanie Strassel, Martha Palmer, Teruko Mitamura, Marjorie Freedman, Heng Ji, et al. A comparison of event representations in deft. In *Proceedings of the Fourth Workshop on Events*, pages 27–36, 2016.

[2] Jan Buys and Phil Blunsom. Oxford at semeval-2017 task 9: Neural amr parsing with pointer-augmented attention. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 914–919, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[3] Simon David Hernandez, Davide Buscaldi, and Thierry Charnois. Lipn at semeval-2017 task 10: Filtering candidate keyphrases from scientific publications with part-of-speech tag sequences to train a sequence labeling model. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 995–999, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[4] Ahmed A Mohamed and Sanguthevar Rajasekaran. Improving query-based summarization using document graphs. In *Signal Processing and Information Technology, 2006 IEEE International Symposium on*, pages 408–410. IEEE, 2006.

[5] Kim Schouten, Flavius Frasincar, and Franciska de Jong. Commit at semeval-2017 task 5: Ontology-based method for sentiment analysis of financial headlines. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 883–887, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[6] Vikram Singh, Sunny Narayan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. Iitp at semeval-2017 task 8 : A supervised approach for rumour evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 497–501, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[7] Wenpeng Yin, Mo Yu, Bing Xiang, Bowen Zhou, and Hinrich Schütze. Simple question answering by attentive convolutional neural network. *arXiv preprint arXiv:1606.03391*, 2016.