Flexibility counts more than precision

Position paper for the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation

David McDonald

BBN Technologies
Cambridge, MA 02138 USA
dmcdonald@bbn.com

Abstract

Today's NLG efforts should be compared against actual human performance, which is fluent and varies randomly and with context. Consequently, evaluations should not be done against a fixed 'gold standard' text, and shared task efforts should not assume that they can stipulate the representation of the source content and still let players generate the diversity of texts that the real world calls for.

1 Minimal competency

The proper point of reference when making an evaluation of the output of a natural language generation (NLG) system is the output of a person. With the exception of the occasional speech error or other predicable disfluencies such as stuttering or restarts, people speak with complete command of their grammar (not to mention their culturally attuned prosodics), and with complete command of their discourse context as it shapes the coherence of what they say and the cohesion of how they say it.

Any NLG system today that does not use pronouns correctly (assuming they use them at all), that does not reduce complex NPs when they describe subsequent references to entities already introduced into the discourse, that does not reduce clauses with common subjects when they are conjoined, or that fails to use any of the other ordinary cohesive techniques available to them in the language they are using is simply not in the running. Human-level fluency is the entrance ticket to any comparative evaluation of NLG systems.

2 Real sources

Similarly, any system that started from a hand-made source representation (as we all did in the 1970s) would not be meeting the minimal standards by which we should measure an NLP system today. Any proposal for a shared evaluation campaign should provide source representations that reflect real data used to do real work for real (preferably commercial) systems.

A good example of a class of real sources is minimally interpreted numerical data sources such as raw instrument readings for weather reports (SumTime) or data points in the movement of stock averages during a day of trading (Kukich 1988). I will propose a more versatile source later.

3 Variation is expected

When I read Winnie-the-Pooh to my daughter at bed time what comes out of my mouth is not always what was in the book, though it always carries the same message. Overworked phases aside, people rarely phrase their content the same way time after time even when they are talking about something they know very well.

This natural level of variation that people exhibit is something that our NLG systems should do as well. It is the only way, for example, that a syn-

thetic character in a computer game that incorporated a proper NLG system would ever be seen as realistic, which is crucial in game-based training systems where suspension of disbelief is required if the training is to be effective.

4 Context is everything

Consider these passages that I clipped from today's news.[1] The first is the title pointing to the full article and was positioned next to a graphic. The second was the small blurb that summarized the content of the article. The third is the equivalent text close to the top of the full article. If we looked at Apple's press release or its quarterly earnings report that prompted this BBC article we would see still different phrasings of this same content.

"Apple profits surge on iPod sales"

"Apple reports a 78% jump in quarterly profits thanks to strong Christmas sales of its iPod digital music player."

"Apple has reported a 78% surge in profits for the three months to 30 December, boosted by strong Christmas sales of its iPod digital music player."

From the point of view of the source representation that a NLG system would use, these three texts are arguably based on the identical content. Some leave out details, others choose different phrasing. What drives the differences is the purpose that the text serves—the context in which is will be used—a flashy title to catch the eye; a short summary; the lead in to a full write up.

5 Where does flexibility come from?

As these example show, a good generator will be sensitive to its context and adapt what it produces accordingly. Still, other than things like freely varying choices of synonyms and semantically neutral variations in linguistic constructions that could be governed by genuinely random 'decisions', most NLG systems prefer to have rationales behind their choices, whether they are the design of the features sets that govern statistical systems or symbolic rules. Where are the rationales for such widely varying surface forms going to come from, and how might they be incorporated in a common data set for evaluation?

I don't believe that we know the answer to this question yet other than that is has something to do with the set and setting deep within the computational entity for whom the generator is working. This calls for research on the kinds of representations that initiate and drive generation and how they encode teleology and psychological motive. No two researchers are likely to agree on what this representation looks like, and for texts like these examples it cannot be reduced to numerical data.

Let me suggest that a clean way to handle this problem is to make the shared data set be *the texts themselves*, with their settings, and to let the players construct whatever representation they want by *parsing* them. Taking the interpretations back far enough to identify a common core content among a set of different texts that are stipulated by a consensus of judges to be conveying essentially the same content should provide some insight into the reason for the difference that just starting from the generation direction would not.

Parsing and regenerating is also a worthy problem in its own right. There is a vast wealth of information that is only available as texts, and DARPA and others are actively developing efforts in 'learning by reading'. I believe that a natural sweet spot for commercial generation work in the future (besides the game world) is in regenerating a common body of content in different genres and with different functions, just as human journalist does after reading a press release. If we can take up this problem collectively as part of a shared task, so much the better.

References

Karen Kukich, 1988. *Fluency in Natural Language Reports*. McDonald & Bolc (eds.) Natural Language Generation Systems. Springer-Verlag series in Symbolic Computing

SumTime: "Generating English Summaries of Time-Series Data. http://www.csd.abdn.ac.uk/~ereiter/sumtime.html

[1] BBN News, 17 January 2007.