Toward Automatic Scoring and Alignment of Narrative Recall

Emily Tucker Prud'hommeaux
Center for Spoken Language Understanding, Oregon Health and Science University

Cognitive and neuropsychological examination protocols often include a narrative recall task in which the subject listens to a brief story and then retells the story to the examiner, either immediately upon hearing the story or after a delay. The examiner scores the retelling by noting how many items from a predetermined list of target words and phrases the subject uses. Examples of this type of examination include the Wechsler Logical Memory task, typically used in diagnosis of neurodegenerative disorders such as Alzheimer's related dementia, and the narrative memory subtest of the NEPSY, an instrument for assessing neuropsychological development in preschool and school-age children.

Typically the scoring is done in real time; that is, the examiner must observe and count the target phrases while the subject is performing the retelling. This can be problematic for a number of reasons. The list of targets can be quite long, making it easy for examiners to lose track of which targets have been used. If the phrases are uttered out of order, they can be missed. In addition, scoring instructions generally allow some degree of paraphrasing, which adds an element of uncertainty and subjectivity.

We propose to enhance the reliability of this procedure by developing an automated scoring system that searches in a retelling for both the targets and their paraphrases generated from aligned parallel corpora. This technique of generating paraphrases will also be used to align retellings to the original narrative, allowing the calculation of more general purpose measures of narrative coherence. The paper presents preliminary results that demonstrate both the difficulties associated with automatic scoring of narrative recall and the potential for improvement using automatic paraphrasing.

One of the battery of psychological and developmental examinations administered to the subjects participating in an ongoing study of autism and developmental language disorder at OHSU is the NEPSY (Korkman et al., 1998). The NEPSY consists of a series of tests that are used to evaluate neuropsychological development in five areas: attention and executive function, language, memory and learning, sensorimotor function, and visuospatial function. Normed on a stratified sample of 1,000 children, the NEPSY is widely used in the clinical community and is standardized relative to a number of other measures such as the Wechsler Intelligence Scale for Children - Fourth Edition (WISC-IV; (Wechlser, 2004)) and the Children's Memory Scale (CMS; (Cohen, 1997)).

> Jim was a boy whose best friend was Pepper. Pepper was a big, black dog. Jim liked to walk in the woods and climb the trees. Near Jim's house was a big oak tree with branches so high that he couldn't reach them. One day Jim decided to climb the tree. He got a ladder from home and carried it to the oak tree. When he got to the top, he looked out over his neighborhood. When Jim started to get down, his foot slipped, his shoe fell off, and the ladder fell to the ground. Pepper sat below the tree and barked. Suddenly, Pepper took Jim's shoe in his mouth and ran off. Jim felt sad. Didn't his friend want to stay with him when he was in trouble? Pepper took the shoe to Anna, Jim's sister. He barked and barked. Finally Anna understood that Jim was in trouble. She took a ladder to the tree and rescued Jim. Wasn't Pepper a smart dog?

Figure 1: Text of NEPSY narrative recall story.

One component of the NEPSY is a narrative recall task, in which the subject listens to and retells a brief story, shown in Figure 1. During the retelling, the examiner notes which items on the NEPSY score sheet's list of target words and phrases are used by the subject. The score for the free recall subsection of the task is the total number of targets used by the subject during the retelling. The subjects analyzed here include 89 children, ages 4 to 8, with performance (non-verbal) IQ great than 70, as measured by the WISC-IV PRI (Wechlser, 2004) for subjects 7 years of age and older and the Wechsler Preschool and Primary Scale of Intelligence PIQ (Wechlser, 2002) for subjects under 7. Each subject's retelling was recorded and manually transcribed at the utterance level. The transcribed retellings were then scored to ensure that no items were overlooked by the examiners and to identify every word or phrase matching one of the targets.

The list of target phrases was expanded by hand to include the substitution of pronouns for proper names; for example, *she helped him* was added as a possible paraphrase of *Anna helped Jim*. The tar-

gets were then located in the retelling transcripts using `grep`. Of the 407 phrases identified by the scorer as matches for targets, 215 were exact matches (52.8%). The remaining phrases identified as matches by the scorer were paraphrases of the targets that included anaphora, synonymy, and syntactic variation.

Generating possible paraphrases by rules, such as replacing content words and proper nouns with pronouns, substituting synonyms for target words, and varying syntactic structure, would require extensive manual annotation and would likely overgenerate paraphrases. Instead, we include in the list of targets actual paraphrases – rather than just theoretically possible paraphrases – using Callison-Burch's (2008) system for generating paraphrases using aligned parallel corpora. Each phrase to be paraphrased is located in the English side of an aligned parallel corpus. The various foreign phrases used to translate that target English phrase are used as pivots to find English paraphrases. The algorithm searches on the foreign side of the corpus for other instances of those phrases and locates the English translations for those phrases. These English phrases are likely to be paraphrases of the original English target phrase.

The Callison-Burch system is pre-trained on the Europarl parallel corpus (Koehn, 2005). Submitting the NEPSY target phrases to the system trained on the Europarl parallel corpora yielded few results, which suggests that a more topical and appropriate parallel corpus is needed in order to generate usable paraphrases. Two in-domain corpora were assembled: 1) a 25,000-sentence English-French parallel corpus of freely available child-oriented texts, including Jules Vernes novels and Hans Christian Andersen's fairy tails (both from Project Gutenberg (Hart, 1997)), and a portion of the Opus corpus of movie subtitles (Tiedemann and Nygaard, 2004); and 2) a 7000-sentence English-English parallel corpus using multiple translations of Grimms' and Andersen's fairy tales. Including the paraphrases generated with these corpora in the list of NEPSY target phrases increased the number of matches found with `grep` from 215 to 222 (54.5%).

The modest gains reported here are likely due to small corpus size and inadequate sentence alignment. Work in progress focuses on expanding the existing corpora and improving the sentence alignment. A more sophisticated sentence alignment algorithm and additional supervision of the alignments, along with the inclusion of larger corpora, are likely to result in more and better paraphrases, which will improve the accuracy of the automated scoring procedure. Although we have found only slight improvements so far using approximate matching, we also expect that allowing approximate matches (e.g., those with some maximum edit distance) may result in more significant improvements with the inclusion of more paraphrases.

Another interesting application of paraphrasing to explore is the use of paraphrases for aligning a retelling with the original narrative in order to measure narrative coherence. The NEPSY task analyzed here is used to measure narrative *recall*, which requires only knowledge of facts, rather than narrative *coherence*, which requires understanding of the order and causes of events and the motivations of protagonists. This distinction is particularly relevant given that the population we study includes children with autism spectrum disorder, who may excel in recalling facts but fail to understand how those facts are related, as well as children with developmental language disorder, who may understand the gist of a story but be unable to express themselves with the exact vocabulary used in the original narrative.

References

Chris Callison-Burch. 2008. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. In Proceedings of EMNLP 2008.

Morris Cohen. 1997. *Children's Memory Scale (CMS).* San Antonio: The Psychological Corporation.

Michael Hart. 1997. Project Gutenberg. `http://www.gutenberg.org`.

Phillip Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation *MT Summit 2005*

Marit Korkman, Ursula Kirk and Sally Kemp. 1998. *NEPSY: A developmental neuropsychological assessment.* San Antonio: The Psychological Corporation.

Jorg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - Parallel and Free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

David Wechsler. 2003. *Wechsler Intelligence Scales for Children - Fourth Edition (WISC-IV).* San Antonio: The Psychological Corporation.

David Wechsler. 2003. *Wechsler Primary and Preschool Scale of Intelligence - Third edition (WPPSI-III).* San Antonio: Harcourt Assessment.