Sentiment Analysis and Lexical Cohesion for the Story Cloze Task

Michael Flor
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08540
mflor@ets.org

Swapna Somasundaran
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08540
ssomasundaran@ets.org

Abstract

We present two NLP components for the Story Cloze Task – dictionary-based sentiment analysis and lexical cohesion. While previous research found no contribution from sentiment analysis to the accuracy on this task, we demonstrate that sentiment is an important aspect. We describe a new approach, using a rule that estimates sentiment congruence in a story. Our sentiment-based system achieves strong results on this task. Our lexical cohesion system achieves accuracy comparable to previously published baseline results. A combination of the two systems achieves better accuracy than published baselines. We argue that sentiment analysis should be considered an integral part of narrative comprehension.

1 Introduction

The Story Cloze Task (SCT) is a novel challenge task in which an automated NLP system has to choose a correct ending for a short story, from two predefined alternatives. This new challenge stems from a long line of research on the types of knowledge that are required for narrative comprehension (Winograd, 1972; Schank and Abelson, 1977). Specifically, it is related to a previous type of challenge, the Narrative Cloze Task (NCT) (Chambers and Jurafsky, 2008).

The SCT departs from the narrow focus of the NCT. It is informed by the interest in the temporal and causal relations that form the intricate fabric of narrative stories. Some previous research on analyzing and learning commonsense information have focused on blogs (Gordon and Swanson, 2009; Manshadi et al., 2008), which are challenging and difficult texts. Other studies have focused on analysis of short fables (Goyal et al., 2013; Elson and McKeown, 2010). Mostafazadeh et al. (2016) produced a large curated corpus of simple commonsense stories, generated via crowdsourcing. Each story consists of exactly five short sentences, with a clear beginning, middle and ending, without embellishments, lengthy introductions and digressions.

For the Story Cloze Task, human authors used four-sentence core stories form the corpus, and provided two different ending sentences - a 'right' one and a 'wrong' one. Some of the 'wrong' endings include logical contradictions, some include events that are impossible or highly unlikely given our standard world knowledge. For example: 1. *Yesterday Stacey was driving to work.* 2. *Unfortunately a large SUV slammed into her.* 3. *Luckily she was alright.* 4. *However her car was destroyed.* Options: 5a. *Stacey got back in her car and drove to work* [wrong]. 5b. *Stacey told the police what happened* [right].

The current SCT has a validation set and a test set, with 1871 stories per set. Each story consists of four sentences, and two competing sentences as story endings. An NLP system is tasked to choose the correct ending from the two alternatives. Systems are evaluated on a simple accuracy measure (number of correct choices divided by number of stories). In this setting, if ending-choices are made randomly, the baseline success rate would be 50%.

In this paper we present our system for the SCT challenge. Section 2 outlines the approach, section 3 describes the algorithms and the results.

2 Approach

Our system is not designed for deep understanding of narrative or for semantic reasoning. The goal of our approach is to investigate the contribution of sentiment and affect in the SCT task. We

Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pages 62–67,
Valencia, Spain, April 3, 2017.

build our system with components that perform sentiment analysis and estimate discourse cohesion. Our system tries to pick the most coherent ending based on the sentiment expectations that the story builds in the minds of the reader. For sentiment we consider positive and negative emotions, feelings, evaluations of the characters, as well as positive and negative situations and events (Wilson and Wiebe, 2005). In cases where sentiment is absent, we rely on lexical cohesion. Our approach is to investigate, if and how, coherence, as modeled using simple methods, can perform in the story completion task.

Consider this story with marked sentiment: 1. *Ron started his new job as a landscaper today* [neutral]. 2. *He loves the outdoors and has always enjoyed working in it* [positive]. 3. *His boss tells him to re-sod the front yard of the mayor's home* [neutral]. 4. *Ron is ecstatic, but does a thorough job and finishes super early* [positive]. Choices for ending (correct option is 5b): 5a. *Ron is immediately fired for insubordination* [negative]. 5b. *His boss commends him for a job well done* [positive].

In this story, there is a positive sequence of events. Hence it would be rather incoherent to have an ending that is starkly negative (incorrect option 5a). If indeed a negative ending were to be applied, it would be a twist in the story and would have to be indicated with a discourse marker to make the story coherent. In short stories, plot twists that relate to sentiment polarity are usually expressed via adverbial phrases, such as 'however', 'unfortunately' or 'luckily', and contrastive connectors, such as 'but' and 'yet'. Notably, some adverbials only indicate contrast to previous context (such as 'however'), while others induce a specific sentiment polarity. For example, 'luckily' indicates positive sentiment, overriding other sentiment-bearing words in the sentence; 'unfortunately' indicates negative sentiment, again overriding other indicators in the sentence. This is seen in the example below where a series of bad (negative) situations suddenly change for the better. The positive twist in the story is indicated by 'luckily'.

Example: 1. *Addie was working at the mall at Hollister when a strange man came in* [negative]. 2. *Before she knew it, Addie looked behind her and saw stolen clothes* [negative]. 3. *Addie got scared and tried to chase the man out* [negative]. 4. *Luckily guards came and arrested him* [overall positive, with an indication for positive story twist]. Ending options (correct is 5b): 5a. *Addie was put in jail for her crime* [negative]. 5b. *Addie was relieved and took deep breaths to calm herself* [positive].

In the absence of sentiment in a story, discourse coherence is, to some extent, captured by lexical cohesion. Take for example the following: 1. *Sam bought a new television.* 2. *It would not turn on.* 3. *He pressed the on button several times.* 4. *Finally Jeb came over to check it out.* Ending options (5b is correct): 5a. *Jeb turned on the microwave.* 5b. *Jeb plugged the television in and it turned on.* Here, even though both sentences introduce a new term ('microwave' and 'plugged'), the latter is semantically closer to the main story.

3 System Description

We construct two systems, one based on sentiment and another based on cohesion. We use the prediction from the sentiment-based system when the story has positive or negative sentiment elements, and back-off to a cohesion-based system when no sentiment is detected.

3.1 Sentiment-based system

Mostafazadeh et al. (2016) presented initial efforts to use sentiment analysis for the SCT. They used two approaches. Sentiment-Full: choose the ending that matches the average sentiment of the context (sentences 1-4). Sentiment-Last: choose the ending that matches the sentiment of the last context sentence. In both cases, they used the sentiment analysis component from the Stanford CoreNLP toolkit (Manning et al., 2014). No details were given on the algorithm they used for the story completion task. These respective models achieved accuracy of 0.489 and 0.514 on the validation set, and 0.492 and 0.522 on the test set.

For our analyses, we used an adapted version of the VADER sentiment dictionary. The original VADER dictionary (Hutto and Gilbert, 2014) contains 7062 lexical entries, with valence (sentiment) scores on the scale from -5 (very negative) to 5 (very positive). We expanded those lexical entries, and added all their inflectional variants, using an in-house English morphological toolkit. Our modified sentiment dictionary has 8255 entries. New words inherited the valence scores of origin words. For all entries, valence scores were rescaled into the range between -1 and 1.

For computing sentiment value for a sentence we filter out stop words (using a list of 250 com-

mon English stopwords) and analyze only content words. For each word, we retrieve its valence from the sentiment dictionary (if present), and sum up the values for the whole sentence. We implement local negation handling - if a sentiment-bearing word is negated (by a preceding word), the sentiment value of the word flips (multiply by -1) (Taboada et al., 2011). In addition we handle twists by checking for adverbials. If a sentence starts with a polarity-inducing adverbial, the sum of polarity values for the story is changed to the sign of the inducing adverbial (positive or negative). For this purpose, we prepared our own dictionary of polarity-inducing adverbials.

The key component in using sentiment scores for SCT is the decision rule: choose the completion sentence whose sentiment score is congruent with the rest of the story. The rule has two parts: a) Choose the completion sentence that has the same sentiment polarity as the preceding story. If the preceding story has positive (negative) sentiment, choose the positive (negative) completion. b) If both completions have same polarity, sign-congruence will not work. In such cases, we choose the completion whose value (magnitude) is closer to the sentiment value of the preceding context.

While analyzing the stories from the validation set with the VADER dictionary, we noted that 78% of the stories have sentiment-bearing words, both in the core sentences (sentences 1-4) and in at least one of the alternative ending sentences. The test set has an even higher percentage of such stories: 86%. The sentiment-based decision rule in the SCT cannot be applied in cases where the core-story or both completion sentences do not have a sentiment value. Thus, in order to test the effectiveness of our sentiment-based approach, we first tested its performance on sentiment-bearing stories only. Results are presented in Table 1. Note that the number of stories-with-sentiment depends on the lexicon. The results clearly indicate that considering the sentiment of the whole preceding context has a strong contribution towards selecting the correct ending (above 60% accuracy). Making a choice while considering the sentiment of only the last context sentence is much less successful (performance is worse than random).

We conducted a similar analysis with another lexicon – MPQA (Wilson et al., 2005), which has only binary valence values. It provided similar (albeit lower) results: 66% of stories in the validation set have sentiment, and accuracy on this set is 0.555. This indicates that even very simplified sentiment analysis has some utility for the SCT.

Set	Sentiment-Full	Sentiment-Last
Validation (1469)	0.679	0.436
Test (1610)	0.607	0.358

Table 1: Sytem accuracy on stories where sentiment is detected (in paretheses: number of stories with sentiment).

3.2 Lexical Cohesion

Our language model for lexical cohesion uses direct word-to-word associations (first-order word co-occurrences). This type of model has been successfully used for analyzing the contribution of lexical cohesion to readability and text difficulty of short reading materials (Flor and Beigman Klebanov, 2014). The current model was trained on the English Gigaword Fourth Edition corpus (Parker et al., 2009), approximately 2.5 billion word tokens. The model stores counts of word co-occurrence collected within paragraphs (rather than windows of set length). We use Positive Normalized PMI as our association measure. Normalized PMI was introduced by (Bouma, 2009); positive NPMI maps all negative values to zero. To calculate lexical cohesion between two sentences (or any other snippets of text), we use the following procedure. First, for each sentence we remove the stopwords. Then, we generate all pairs of words (so that one word comes from first sentence and the other word comes from the second sentence), retrieve their association values from the model, and sum up the values. The sum of pairwise associations can be used as a similarity (or relatedness) measure. We also experimented with average (dividing the sum by the number of pairs), but the sum performed slightly better in our experiments. For the SCT task, for each story, we computed lexical cohesion between sentences 1-4, taken together as a paragraph, and each of the competing completion sentences (LexCohesion Full). The decision rule is to choose the completion sentence that is more strongly associated with the preceding story. Accuracy is 0.534 on the validation set and 0.527 on the test set (with 1871 stories in each set). We also computed lexical cohesion between the last sentence of context and each of the competing endings (LexCohesion

Last). For this condition, accuracy is 0.556 on the validation set and 0.536 on the test set. Our results are comparable to those of (Mostafazadeh et al., 2016), who used vector-space embeddings and obtained 0.545 and 0.536 on the validation set, 0.539 and 0.522 on the test set.

3.3 Combining the Models

To provide a full algorithm for the SCT task, we use our sentiment-based algorithm, and only back off to our lexical cohesion model when sentiment is not detected in the story (sentences 1-4) or in neither of the ending sentences. Results are presented in Table 2. Best accuracy is achieved by combining the Sentiment-Full model with LexCohesion-Last: 0.654 on the validation set and 0.620 on the test set. These results outperform the previously published best baseline of 0.604 and 0.585 (Mostafazadeh et al., 2016).

Set	Sentiment-Full + LexCohesion-Full	Sentiment-Full + LexCohesion-Last
Validation	0.639	0.654
Test	0.618	0.620

Table 2: System accuracy on all stories in each set.

4 Discussion

The role of affect and emotion has long been noted for human story comprehension (Kneepkens and Zwaan, 1994; Miall, 1989) and in AI research on narratives (Lehnert and Vine, 1987; Dyer, 1983). Stories are typically about characters facing conflict. Sometimes the plot complications (negative events or situations) have to be overcome. In many such cases, one expects to encounter sentiment expressions in the story. Not surprisingly, we found that a large proportion of the stories, in both validation and test sets, have sentiment-bearing words. Thus, it is only natural to expect that sentiment analysis should be able to impact SCT. Our approach is to look at the polarity of sentiments and for sentiment congruence in a story. Using a sentiment dictionary that assigns sentiment values on a continuous scale, and looking only at lexically indicated sentiment, our algorithm chooses the correct ending in more than 60% of stories (when sentiment is detected).

We have demonstrated that even a rather simple, lexically-based sentiment analysis, can provide a considerable contribution to accuracy in the SCT. Our system only evaluates the congruence of two competing solutions, without attempting to develop a deep understanding of the story. For example, our system does not have the capability to reason that a car that has just been wrecked cannot be used to drive away. However, we consider that analyzing the sentiment of a story is not a shallow task (even if it is technically rather simple). We believe that human-level understanding of narrative involves many facets, including chains and schemas of events, plot units, character goals and states, etc. Handling each of them presents unique challenges to an NLP system. We argue that the sentiment aspect of narratives is one of the key aspects of stories. In fact, sentiment is a very deep aspect of narrative (Mar et al., 2011), one that we have only begun to explore. As the SCT focuses on very short stories, it is interesting to note that patterning of sentiment and affect has also been shown to exist on the scale of long novels (Reagan et al., 2016).

While our dictionary-based sentiment analysis was quite successful, we note that it should be viewed only as a starting point for investigating the role of sentiment in narrative comprehension. In the SCT, there are cases were dictionary-based sentiment detection fails to detect the sentiment in a story. For example: 1. *Brad went to the beach.* 2. *He made a sand castle.* 3. *He jumped into the ocean waters.* 4. *He swam with the small fish.* Options: 5a: *Brad's day went very badly.* 5b: *Brad then went home after a nice day.*

The above story has quite positive connotations, but none of the lexical terms from sentences 1-4 carry sentiment values in the dictionary. Our system detects sentiment in each of the competing ending sentences, but since no sentiment was detected in sentences 1-4, choice of the ending is relegated to lexical cohesion, rather than sentiment. A human reader would choose the second ending, based on positive sentiment connoted by the events.

5 Conclusion

In this paper we described a simple approach that combines sentiment- and cohesion-based systems for the Story Cloze Task. While previous research found sentiment analysis to be ineffective on this task, we proposed a new approach, using a rule that estimates sentiment congruence in a story. Our system achieves accuracy of 0.654 on the validation set and 0.620 on test set, mostly due to the

strong contribution from sentiment analysis. Our results provide support to the notion that sentiment is an important aspect of narrative comprehension, and that sentiment-analysis can be a strong contributing factor for NLP analysis of stories. There are a number of avenues for further exploration, such as using machine learning methods to combine different types of information that go into making a story, using vector spaces, automated reasoning and extending the feature set to capture other aspects of language understanding.

References

Gerloff Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference*, pages 31–40. Gunter Narr Verlag, Tubingen.

Nathanael Chambers and Daniel Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of the 46th Annual Meeting of the ACL*, pages 789–797, Columbus, OH, USA, June. Association for Computational Linguistics, Association for Computational Linguistics.

Michael G. Dyer. 1983. The role of affect in narratives. *Cognitive Science*, 7:211–242.

David K. Elson and Kathleen R. McKeown. 2010. Building a bank of semantically encoded narratives. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Michael Flor and Beata Beigman Klebanov. 2014. Associative lexical cohesion as a factor in text complexity. *International Journal of Applied Linguistics*, 165(2):223–258.

Andrew S. Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop*, San Jose, CA, USA.

Amit Goyal, Ellen Riloff, and Hal Daume III. 2013. A computational model for plot units. *Computational Intelligence*, 3(29):466–488.

C.J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of The 8th International AAAI Conference on Weblogs and Social Media (ICWSM14)*, Ann Arbor, MI, USA. Association for the Advancement of Artificial Intelligence.

E.W.E.M Kneepkens and Rolf A. Zwaan. 1994. Emotions and literary text comprehension. *Poetics*, 23:125–138.

Wendy G. Lehnert and Elaine W. Vine. 1987. The role of affect in narrative structure. *Cognition and Emotion*, 1(3):299–322.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60. Association for Computational Linguistics.

Mehdi Manshadi, Reid Swanson, and Andrew S. Gordon. 2008. Learning a probabilistic model of event sequences from internet weblog stories. In *Proceedings of 21st Conference of the Florida AI Society, Applied Natural Language Processing Track*, Coconut Grove, FL, USA.

Raymond A. Mar, Keith Oatley, Maja Djikic, and Justin Mullin. 2011. Emotion and narrative fiction: Interactive influences before, during, and after reading. *Cognition and Emotion*, 25(2):818–833.

David S. Miall. 1989. Affective comprehension of literary narratives. *Cognition and Emotion*, 3(1):55–78.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Pushmeet Kohli Lucy Vanderwende, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English gigaword fourth edition. https://catalog.ldc.upenn.edu/LDC2009T13.

Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(31).

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum, Hillsdale, NJ, USA.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-basedmethods for sentiment analysis. *Computational Linguistics*, 2(37):267–307.

Theresa Wilson and Janyce Wiebe. 2005. Annotating attributions and private states. In *Proceedings of ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 53–60, Ann Arbor, MI, USA. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pages 347–354, Vancouver. Association for Computational Linguistics.

Terry Winograd. 1972. *Understanding Natural Language*. Academic Press, Inc., Orlando, FL, USA.