BALLGAME: A Corpus for Computational Semantics

Ezra Keshet, Terry Szymanski, and Stephen Tyndall

University of Michigan
E-mail: {ekeshet,tdszyman,styndall}@umich.edu

Abstract

In this paper, we describe the Baseball Announcers' Language Linked with General Annotation of Meaningful Events (BALLGAME) project – a text corpus for research in computional semantics. We collected pitch-by-pitch event data for a sample of baseball games and used this data to build an annotated corpus composed of transcripts of radio broadcasts of these games. Our annotation links text from the broadcast to events in a formal representation of the semantics of the baseball game. We describe our corpus model, the annotation tool used to create the corpus, and conclude by discussing applications of this corpus in semantics research and natural language processing.

1 Introduction

The use of large annotated corpora and treebanks has led to many fruitful research programs in computational linguistics. At the time of this writing, Marcus et al. (1993), which introduces the University of Pennsylvania Treebank,[1] has been cited by over 3000 subsequent papers.[2] Such treebanks are invaluable for the training and testing of large-scale syntactic parsers and numerous other applications in the field of Computational Syntax.

Unfortunately for the field of Computational Semantics, there are few corresponding annotated corpora or treebanks representing the formalized meaning of natural language sentences, mainly because there is very little agreement on what such a representation of meaning would look like for arbitrary text. To overcome this obstacle, several recent studies have turned to the arena of sports, pairing natural language with game statistics in several domains, including RoboCup soccer (Liang et al., 2009; Chen et al., 2010), soccer (Theune and Klabbers, 1998; Saggion et al., 2003), American football (Barzilay and Lapata, 2005; Liang et al., 2009), and baseball (Fleischman, 2007).

We have adapted this approach in the creation of a semantics-oriented corpus, using the domain of major-league baseball. The information state of a baseball game can be represented with a small number of variables, such as who is on which base, who is batting, who is playing each position, and the current score and inning. There is even a standard way of representing updates to this information state.[3] This makes baseball a logical stepping stone to a fuller representation of the world. We also chose baseball for this corpus because of the volume of data available, in the form of both natural language descriptions of events and language-independent game statistics. Most of professional baseball's thousands of games per year have at least two television broadcasts (home and away) and at least two radio broadcasts, often in multiple languages. The scorecard statistics for each game are also kept and made available on the internet, along with complete ordered lists of in-game events. These resources, coupled with a high-coverage syntactic parser, allow one to link natural language utterances with representations of their syntax and semantics.

[1] http://www.cis.upenn.edu/~treebank/

[2] http://scholar.google.com/scholar?cites=7124559111460341353

[3] See example scorecards at http://swingleydev.com/baseball/tutorial.php.

2 Corpus Design

The basic design of the BALLGAME corpus is a mapping between *spans* of text and *events* in a baseball game. The raw text comes from the transcribed speech of announcers broadcasting the radio play-by-play of a professional baseball game. This text is chunked into spans, and these spans are then labeled according to the following scheme:

- *Event* is the label given to a span that describes an event in our representation of the game for the first time. (Examples of events are simultaneous descriptions of pitches, plays, and stolen bases.)
- *Recap* is the label given to a span that correlates with prior events in the game. (Examples of recaps are when the announcer states the current score or strike count, or summarizes the current batter's previous at-bats.)
- *Banter* is the label given to a span that does not relate to an event in the game. The majority of spans are labeled as banter. (Examples of banters are "color" commentary, any discussion of the day's news, other baseball games, advertisements, etc.)

The term "span" has no linguistic significance, although spans often turn out to be sentences or clauses. Each span from the text that is labeled as an event is linked to one or more events in the model of the game as shown in Figure 1. Not every event is linked to a span of text, since some events go unmentioned by the announcers.

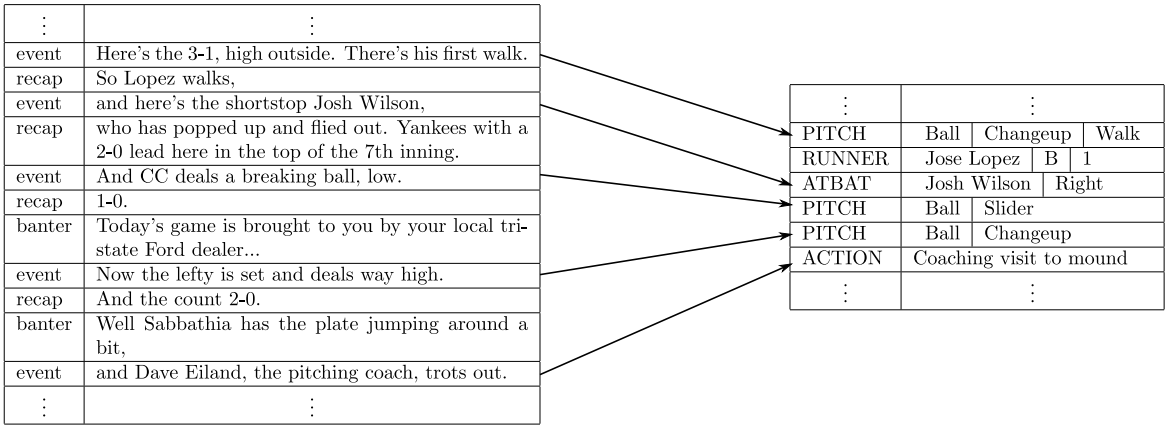

Figure 1: Illustration of a portion of the corpus: event spans of the text (on the left) are associated with events from a standardized description of the ballgame (on the right).

We model each game as a time-ordered sequence of baseball events, designed so that the state of the game at any given point, while not explicitly represented, can be computed given the events from the start of the game up to that point. We use a simple event model inspired by the comprehensive scoring system developed by Retrosheet,[4] but modified to match our needs and data resources. For example, most baseball scoring systems are at-bat-based, but this system is too coarse-grained for our purposes. Therefore, we use a system in which the fundamental event type is the pitch. Every baseball action from the start of the pitcher's motion until the end of the play (a hit or an out) is categorized as a PITCH event. Several other event types exist to accommodate other plays (e.g. balks, pick-offs), non-play actions (e.g. coaching visits to the mound, rain delays), and procedural activities (e.g. ejections, player substitutions).

In addition to a category, each event has multiple attribute values. The possible attributes depend on the category. A PITCH event, for example, has attributes describing the type, speed, and location of the pitch as well as whether it results in a ball, strike, play, etc. If the result is a play, then there are additional

[4] http://www.retrosheet.org

attributes describing the fielders involved in the defensive play. On the other hand, a PICKOFF event has different attributes, describing which base the ball was thrown to, whether it resulted in an out, etc.

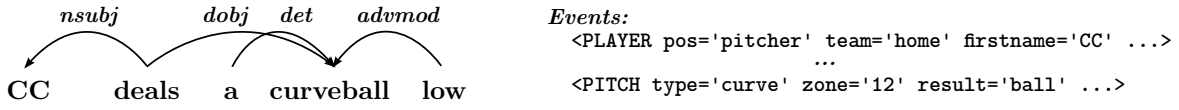

Figure 2: Example of a dependency parsed transcript line and corresponding events.

In the future, we plan to add syntactic parse information for each span such as that generated using the Stanford Parser (De Marneffe et al., 2006). Using an explicit syntactic representation, like the one illustrated in figure 2, it will be possible to label more detailed correlations between the text and the meaning. Even without explicit annotation, statistical learning methods could be used to infer, e.g., that the word "curveball" in the sentence in figure 2 correlates with the semantic attribute `type='curve'`, or that the word "CC" correlates with a specific PLAYER entity. While the annotations in the corpus exist only at the sentence or phrase level, this type of further processing could push the annotation down to the word level, facilitating the study of lexical semantics and semantic transformations of syntactic structures.

3 Corpus Creation

Student transcribers use a custom-created transcription and annotation tool, illustrated in Figure 3, to add data to the corpus. They listen to and transcribe the radio broadcast, while simultaneously chunking the text into spans as described above. Each span is labeled *banter, event*, or *recap*, and, if the span describes an *event*, the student selects the corresponding event(s) from the event column.

Annotators have access to a style guide to encourage consistency. This guide sets out two main principles: first, the transcript of an inning, taken as a whole, should be read like a well-edited, consistently formatted document; and second, all and only the events explicitly mentioned by the radio announcers should be linked to events in the game model.

Although spans are displayed as separate lines in the transcription tool, in order to maintain this first style principle, we ask the students to imagine that all spans of the transcript are pasted together in sequence to form a normal transcript of the game. Thus, they are asked not to put ellipses or dashes at the end of spans nor to capitalize the beginnings of spans that do not begin sentences. Also included in this principle is a standardized formatting style for baseball statistics, such as strike counts, scores, and batting averages, so that, for instance, "the count is two and oh" is transcribed "the count is 2-0".

The second principle set out in the annotation style guide is meant to ensure that the events linked to a particular utterance are as close as possible to the "meaning" of that utterance. Integral to this process is consistently distinguishing the categories of *event*, *recap* and *banter*. Since recap and banter spans do not relate to events in the model, it is important to keep them separate from the event spans to get the most accurate data. Even given the descriptions of these categories from section 2, ambiguous cases still do arise on occasion. For instance, one common difficulty is distinguishing *event* from *recap* when an announcer discusses a play immediately after it happens. In such cases, in keeping with our annotation principle, we use the rule of thumb that only new information is annotated as *event*; old information is *recap*. We also adopt the rule that only game events that are explicitly stated by the announcer should be linked to spans; for example, if the announcer merely states the name of the batter (e.g. "Cust takes a first-pitch strike") in the process of describing the first pitch of his at-bat, then this should not reference the ATBAT event that indicates the arrival of a new batter at the plate. On the other hand, an explicit mention (e.g. "Here's Cust.") should.

In the final steps of the annotation process, each transcript is reviewed and corrected by a second annotator to reduce errors and further promote consistency across annotators.

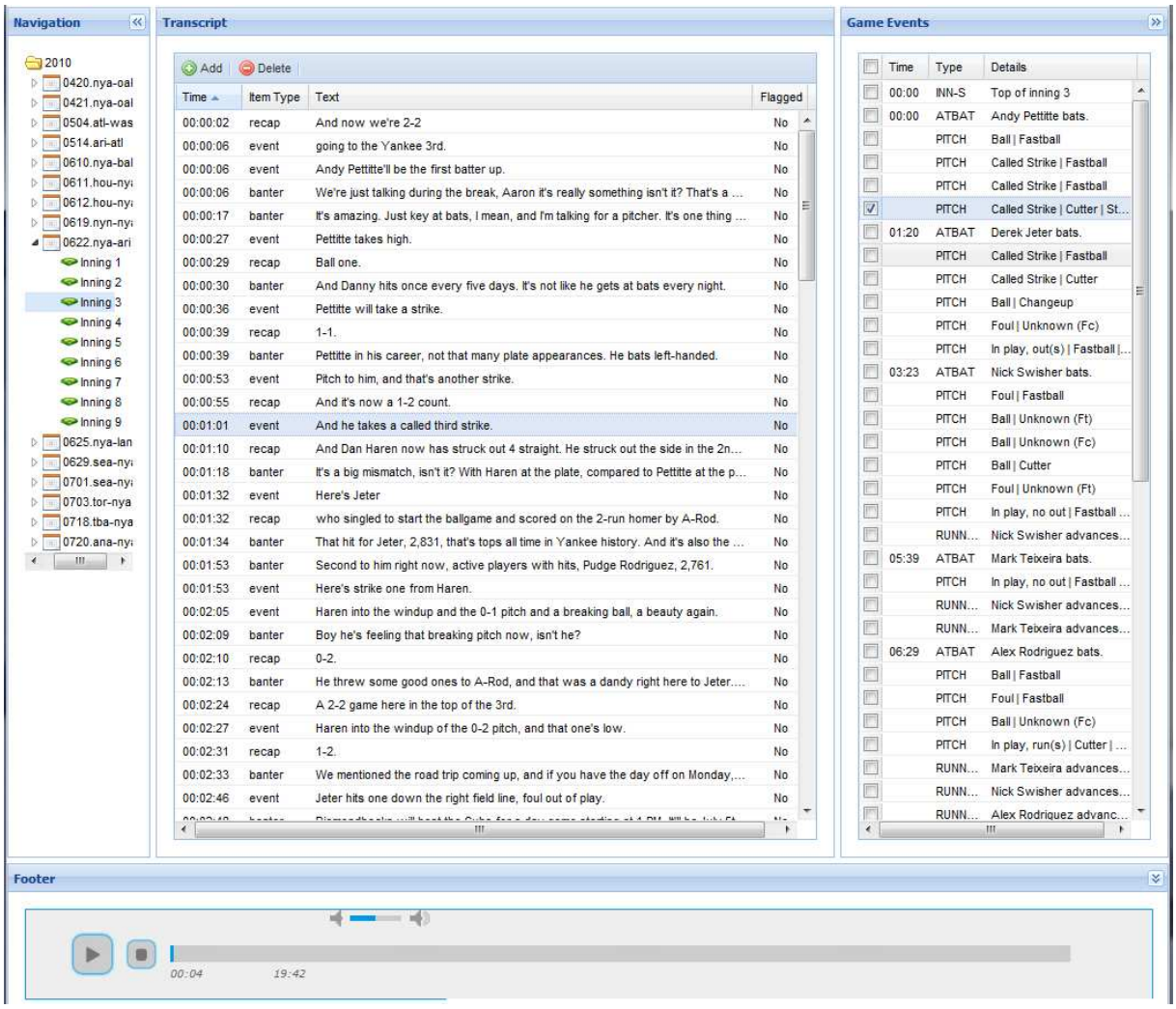

Figure 3: Screen shot of online annotation tool.

4 Potential Applications

Since this corpus links natural language utterances with complete semantic representations which fully describe the state of the baseball game, it has a number of applications for research in computational semantics. While the domain is limited, and the "meaning" of a baseball game does not approach the complexity of the possible "meanings" in the real world, nevertheless this corpus should be a useful resource both for developing NLP tools and for studying theories of language and meaning.

One application domain for this type of data is natural language generation and understanding, and much prior work connecting sports commentaries to statistics or events falls into this domain. One related generation task is to generate textual summaries of complete games: Theune and Klabbers (1998) generated spoken Dutch summaries of soccer matches, and Barzilay and Lapata (2005) investigate the relationship between textual NFL recaps and the box scores of the games. More similar to our project is the RoboCup announcer system of Chen et al. (2010), which produces play-by-play commentary (in English and Korean) of simulated RoboCup soccer matches. Our corpus could certainly be used to train systems that predict the event structure given the text of the commentary, or vice-versa.

In the domain of information extraction, our corpus could be used to train systems to infer representations of meaning from texts. In many domains, the same word or phrase can appear in a variety of different contexts with different ramifications. For example, the phrase "home run" in a baseball commentary may mean that a home run has just occurred, or it may refer to a home run in a previous game, or a player's home-run totals for the season, etc.. Fleischman (2007), using a collection of video

broadcasts of baseball games, combines natural language processing with artificial vision technology to resolve when events like home runs actually occur, in order to facilitate retrieval of relevant video clips. Using our corpus, one could design a system to perform the same task based purely on the textual data, perhaps to extend this same task to radio broadcasts as well as television broadcasts. Given the corpus labels of *event*, *recap*, and *banter*, a classifier could be built to identify only the *event* regions, and an extraction system could identify the relevant semantic features (e.g. player names, types of events).

While generation and understanding are tasks most applicable to this corpus, we hope researchers will find additional innovative uses of the corpus. For example, given that we plan to incorporate a number of baseball games with commentary both in English and Spanish, there is a potential connection to machine translation, particularly approaches that utilize comparable (rather than parallel) corpora. In our corpus, the comparable sections (i.e. the *event*-labeled regions) are explicitly aligned with one another, which is not usually the case in comparable corpora. Also, the corpus could prove useful for research on formal semantics, despite the fact the meaning representation is not particularly rich compared to modern semantic theory, and the jargon and speech styles are very specific to the domain of baseball sportscasts.

5 Conclusion

We have presented an overview of the BALLGAME annotated corpus for research in computational semantics, as well as a description of our procedure for annotation and the specialized annotation tool we developed for this purpose. To date, the corpus contains sixteen three- to four-hour-long major league baseball radio broadcasts, transcribed and annotated as described above. This represents 237,100 transcribed words in 13,382 spans (6,511 *banter*; 3,994 *event*; 2,877 *recap*). Work is ongoing, and the goal is to complete fifty games by the end of the year. We believe this corpus, by pairing natural language text with formalized representations of meaning, will prove useful for many types of NLP research.

References

Barzilay, R. and M. Lapata (2005). Collective content selection for concept-to-text generation. In *Proceedings of HLT/EMNLP*, pp. 331–338.

Chen, D., J. Kim, and R. Mooney (2010). Training a multilingual sportscaster: Using perceptual context to learn language. *Journal of Artificial Intelligence Research 37*(1), 397–436.

De Marneffe, M., B. MacCartney, and C. Manning (2006). Generating typed dependency parses from phrase structure parses. In *LREC 2006*.

Fleischman, M. (2007). Situated models of meaning for sports video retrieval. In *NAACL-HLT 2007*, pp. 37–40.

Liang, P., M. Jordan, and D. Klein (2009). Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, pp. 91–99.

Marcus, M., B. Santorini, and M. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics 19*(2), 313–330.

Saggion, H., J. Kuper, H. Cunningham, T. Declerck, P. Wittenburg, M. Puts, E. Hoenkamp, F. de Jong, and Y. Wilks (2003). Event-coreference across multiple, multi-lingual sources in the Mumis project. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics: Volume 2*, pp. 239–242.

Theune, M. and E. Klabbers (1998). GoalGetter: Generation of spoken soccer reports. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, pp. 292–295.