Multi-Resolution Language Grounding with Weak Supervision

R. Koncel-Kedziorski, Hannaneh Hajishirzi, and Ali Farhadi
University of Washington
{kedzior,hannaneh,farhadi}@washington.edu

Abstract

Language is given meaning through its correspondence with a world representation. This correspondence can be at multiple levels of granularity or *resolutions*. In this paper, we introduce an approach to multi-resolution language grounding in the extremely challenging domain of professional soccer commentaries. We define and optimize a factored objective function that allows us to leverage discourse structure and the compositional nature of both language and game events. We show that finer resolution grounding helps coarser resolution grounding, and vice versa. Our method results in an F1 improvement of more than 48% versus the previous state of the art for fine-resolution grounding[1].

1 Introduction

Language is inextricable from its context. A human language user interprets an utterance in the context of, among other things, their perception of the world. Grounded language acquisition algorithms imitate this setup: language is given meaning through its correspondence with a rich world representation. A solution to the acquisition problem must resolve several ambiguities: the segmentation of the text into meaningful units (spans of words that refer to events); determining which events are being referenced; and finding the proper alignment of events to these units.

Historically, language grounding was only possible over simple controlled domains and rigidly structured language. Current research in grounded language acquisition is moving into real-world environments (Yu and Siskind, 2013). Grounding sports commentaries in game events is a specific instance of this problem that has attracted attention (Liang et al., 2009; Snyder and Barzilay, 2007; Hajishirzi et al., 2012), in part because of the complexity of both the language and the world representation involved.

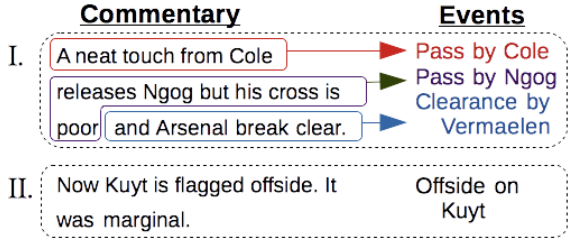

Figure 1: An example of the multiple resolutions at which soccer commentaries refer to events: The utterance level alignments are shown in the black dashed boxes. The first utterance can be further broken into the fragment-level alignments shown; the second cannot be decomposed further.

The language employed in soccer commentaries is difficult to ground due to its dense information structure, novel vocabulary and word senses, and colorful, non-traditional syntax. These challenges conspire to foil most language processing techniques including automated parsers and word-sense disambiguation systems.

In addition to the structural problems presented by the language of soccer commentaries, the problem of reference is further complicated by the fact that for game events (and other real-world phenomena) there is no standardized meaningful linguistic unit. Utterances ranging from a single word to multiple sentences can be used to refer to a single event. For example, in Figure 1 the first four words of commentary (I) refer to a single event, as does the entirety of (II).

[1]Source code and data are available at http://ssli.ee.washington.edu/tial/projects/multires/

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 386–396,
October 25-29, 2014, Doha, Qatar.

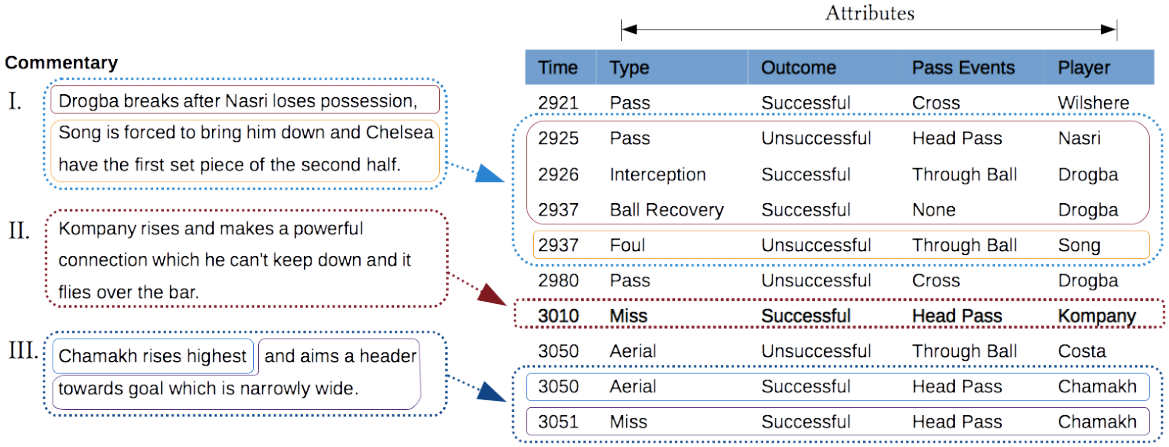

Figure 2: An example of the different levels of granularity present in the soccer data. The dashed boxes on the left denote *utterances* made by the commentators. Solid boxes denote fragments that cannot be decomposed into finer resolution alignments. The table on the right is a portion of the detailed listing of game events.

Turning our attention to Figure 2, sometimes a fragment refers to a combination of events and no further decomposition is available, such as the first fragment of commentary (I). Moreover, it is sometimes desirable to construct a complex of events by determining all the events corresponding to a particular collection of words. For instance, we would want to be able to align the whole of (I) with all the events in the corresponding dashed box. This suggests studying language grounding at multiple levels of granularity (resolutions).

We use *resolution* to describe the continuum of meaningful units which exist in human language[2]. These resolutions interact in a complicated way, with clues from different resolutions sometimes combining to produce an effect and sometimes negating one another. With enough training data, one could hope to learn the details of the interactions of various resolutions. However, the expense of producing or obtaining supervised training data at multiple resolutions is prohibitive.

To address all these complications, we introduce weakly-supervised multi-resolution language grounding. Our method makes use of a factorized objective function which allows us to model the complex interplay of resolutions. Our language model takes advantage of the discourse structure of the commentaries, making it robust enough to handle the unique language of the soccer domain. Finally, our method relies only on loose temporal co-occurrence of events and utterances as supervision and does not require expensive annotated training data.

To test our method we augment the Professional Soccer Commentary Dataset (Hajishirzi et al., 2012) with fragment-level event alignment annotations. This dataset is composed of commentaries for soccer matches paired with event logs produced by Opta Sportsdata and includes human annotated gold alignments[3]. We achieve an F1 improvement of over 48% on fragment-level alignment versus a previous state-of-the-art. We are also able to leverage the interplay of fragment- and utterance- level alignments to improve the previous state-of-the-art utterance-alignment system.

2 Challenges

Syntactic Limitations: Syntax is used to structure the information provided by an utterance, and so it seems intuitive that syntactic relations could be leveraged in this task. For example, consider utterance (III) in Figure 2. The multi-resolution grounding of (III) would provide a *segmentation* of the utterance – or a division of the utterance into the fragments which refer to separate events. In (III), there is an obvious syntactic correlate to the correct segmentation: each verb phrase within the conjunction headed by "and" identifies a separate event. Parsing (III) to an event-based semantics like that of Davidson (1967), one could associate each verb in an utterance with a game event and achieve the desired segmentation.

[2]Though it is tempting to discritize meaning in text, Chafe (1988) shows that readers imbue text with meaningful intonational patterns drawn from the potentially continuous space of auditory signals.

[3]Our updated dataset is available at `http://ssli.ee.washington.edu/tial/projects/multires/`

Unfortunately, there is a preponderance of examples such as (II) in Figure 2, where 4 verbs are used to describe a single "miss" event. (II) illustrates just one of the many difficulties of using syntactic information – elsewhere, events are referenced without an explicit verb whatsoever (such as the use of the phrase "into the books" to refer to a foul event). What is needed instead is a language model that is powerful enough to proscribe some structure yet robust enough to allow the world representation to determine which pieces of language are referring to which referent or set of referents.

Complex Interplay between Resolutions: Language refers at a variety of resolutions, and the relationship between nested reference scopes is complex. A single or few words can indicate entities or properties; full phrases are often needed to denote an action; complex events like a missed shot may take up to several phrases of narration to properly describe. A soccer commentator does not encode every detail necessary for proper alignment and segmentation into their utterances, but rather only enough to make clear to another with similar world knowledge what is meant. A language grounding method is at a severe disadvantage when faced with such implicit information.

Instead, a successful method can make heavy use of the limited lexical, phrasal, and discourse structural cues provided in an utterance, as the different resolutions rely on these different contextual clues to meaning. At finer resolutions one can rely more on the lexical meanings of the words; at medium resolutions, compositionality can be leveraged; at coarser resolutions, discourse features come into play. These cues interact in a complicated way, providing additional challenge.

Consider again Figure 2. In (III), the temporal discourse marker "and" marks the division between the fragments referring to each event. In (I) the same word (used again as a temporal discourse marker) is used to elaborate on the single "foul" event being described in the second fragment. A human (with sufficient understanding of soccer) knows that, despite being separated by the discourse marker, the phrases "bring him down" and "set piece" both refer to the foul. A language grounding algorithm that can model the interaction between such word-level and utterance-level cues can successfully segment both (I) and (III).

Supervision: For language grounding generally, and multi-resolution grounding specifically, supervised training data is expensive to produce. Also, the various grounding domains of interest are highly independent of one another (Liang et al., 2009). In the face of these issues, the ideal correspondence between language and world representation would be learned with as little supervision as possible.

3 Problem Definition

We define the problem of multi-resolution language grounding as follows: Given a temporal evolution of a world state (a sequence of events) and an overlapping natural language text (a sequence of utterances), we want to learn the best correspondences between the language and the world at different levels of granularity (Figure 2).

To set up notations, for each utterance represented as a set of words $W = \{w_1, w_2, \ldots, w_n\}$, we want a segmentation which expresses the relationship of the words to the events which they describe.

Let $\mathcal{S}$ denote a set of all possible segmentations of W. Then $\mathcal{S} = \{S | S \text{ is a segmentation of } W\}$. A segmentation S is in turn a set of non-overlapping fragments ($S = \{s_i\}$), where each fragment is a consecutive sequence of words from the utterance W. For example, for utterance (III) from Figure 2, one possible (incorrect) segmentation is $S = \{s_1, s_2, s_3\}$ for $s_1 = \{$*Chamakh rises highest*$\}$, $s_2 = \{$*and aims a header*$\}$, and $s_3 = \{$*towards goal which is narrowly wide*$\}$.

An alignment consists of a segmentation S and a mapping E from fragments of S to the set of all events $\mathbb{E}$. For example, the segmentation S could be mapped as $E = \{\langle s_1, e_2\rangle, \langle s_2, e_3\rangle, \langle s_3, e_1\rangle\}$, with e_1 being an Aerial Challenge, e_2 being a missed attempt on goal, and e_3 being an out of bounds penalty. Let $\mathcal{E} = \mathcal{S} \times \mathbb{E}$ denote the set of all possible alignments.

As we show in Figure 2, events are composed of the various attributes *Time, Type, Pass Events, Outcome*, and *Player*. For example, the aerial event in Figure 2 has the attributes and values *type:aerial, outcome:successful, pass events:head pass,* and *player:Chamakh*.

Finally, we denote the values for the attributes of each e_j as e_j^a, where a ranges over the different attributes of events as represented in the data.

We define the multi-resolution grounding of W

into $\mathbb{E}$ as the best segmentation S and alignment E that maximize the joint probability distribution:

$$\arg\max_{S \in \mathcal{S}, E \in \mathcal{E}} P(S, E|W) \quad (1)$$

This optimization[4] can be accomplished through the use of supervised learning. However, training data is expensive and tedious to produce for the grounding problem, especially at multiple resolutions. Additionally, the complexity of the language in this domain would result in very sparse associations.

Yet if we knew some of the correct fine-resolution alignments, we could use that information to produce good coarse resolution alignments, and vice versa. Therefore, we formulate a factorized form of the above objective which allows us to learn features specific to aligning at the utterance, fragment, and attribute resolutions. Our method can be optimized with only weak supervision (loose temporal alignments between utterances and a set of events occurring within a window of the utterance time).

We can evaluate such a correspondence in several ways. For each utterance, can we predict the correct events to which this utterance refers? This is the problem of utterance-level alignment.

We can also evaluate based on events: for each event, can we identify the minimal text span(s) which refers to this event? We want a tight correspondence because loose, overlapping alignments are not semantically satisfying. However, we do not want to under associate: human language makes reference at a variety of levels (the word level, the phrase level, the utterance level, and beyond). It is important to correctly identify all and only the words which correspond to a given event. This is the fragment-level alignment problem. We show that good fragment-level alignments will improve utterance-level alignment, and vice versa.

Since events are composed of their attributes, we can imagine a very fine resolution grounding of individual words to individual attributes. In fact, our solution involves producing such a grounding and composing the fragment- and utterance-level alignments therefrom.

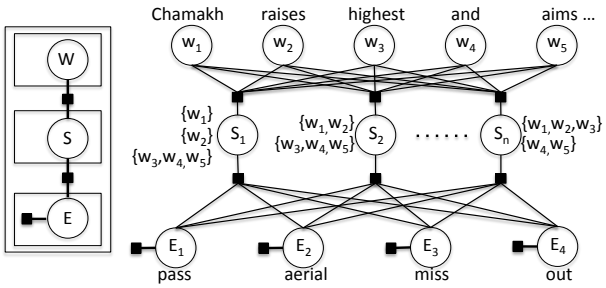

Figure 3: Factor graph for $P(S, E|W)$. Here the w_i are the words of utterance W, S_j are the possible segmentations of W, and E_k are different events.

4 Our Method

We have formulated the grounding problem as an optimization of the joint probability distribution $P(S, E|W)$, which returns the best segmentation and accompanying event alignments given an utterance W. Optimizing this function in the domain of real world language, however, is a difficult problem. Utterances are long here, and there are many events which could be grounded to each. Furthermore, the cardinality of the set of possible segmentations is combinatorially large.

Therefore we decompose Equation 1 using the factor graph depicted in Figure 3. We write the joint probability distribution as a product of the following two potential functions:

$$P(S, E|W) \stackrel{\text{def}}{=} \frac{1}{Z} \prod_{s \in S} \Psi^{\text{align}}(E, s) * \Phi^{\text{seg}}(s, W) \quad (2)$$

where Ψ^{align} is a function for scoring the alignment E for fragment s and Φ^{seg} scores how good a fragment s is for the utterance W, and Z is for normalization.

To optimize Equation 2 it is not practical to search the space of possible S, E combinations (this space is combinatorially large). However, we can optimize the factored form using dynamic programming. We first describe how to find values for each of the potentials in sections 4.1 and 4.2. In section 4.3 we describe the dynamic programming approach to optimization.

4.1 Event Alignments Given Segmentation

The potential function $\Psi^{\text{align}}(E, s)$ takes as inputs a fragment s from segmentation S and a candidate alignment E for S and returns a score for E with

[4] As this and future equations are conditioned on the set of all events $\mathbb{E}$, we omit this variable from the equations for notational simplicity.

regards to s. It is here that we produce the multi-resolution alignments; s can vary in size from a single word to a whole utterance. ψ^{align} decomposes as the following:

$$\Psi^{\text{align}}(E, s) = \Psi^{\text{prior}}(E) * \Psi^{\text{affinity}}(s, E) \quad (3)$$

where the priors (Ψ^{prior}) are confidence scores for an alignment E with the whole utterance as given by Hajishirzi et al. (2012), which fits an exemplar SVM to each utterance/event pair. An exemplar SVM is an SVM fit with one positive and many negative instances, allowing us to define an example by what it is not (Malisiewicz et al., 2011; Shrivastava et al., 2011).

Ψ^{affinity} scores the affinity between a fragment s and the event e_j to which it is aligned. We use the term affinity as a measure of the goodness of an alignment. Intuitively, a fragment s will have a higher affinity for an event e_j if s describes that event well. Formally, the affinity between s and e_j amounts to a product of the affinity between each word $w_i \in s$ and e_j. Since e_j is defined by a collection of attributes, we can compose a score for w_i with e_j from the affinity between w_i and each attribute a of e_j.

$$\begin{aligned}\Psi^{\text{affinity}}(s, E) &= \prod_{w_i \in s, e_j \in E} \psi^{\text{atr.}}(w_i, e_j) \\ &= \prod_{w_i \in s, e_j \in E} \max_a \psi(w_i, e_j^a) \quad (4)\end{aligned}$$

where e_j is the event to which s is aligned in alignment E, $\psi^{\text{atr.}}(w_i, e_j)$ is the affinity between w_i and event e_j, and $\psi(w_i, e_j^a)$ is the affinity between w_i and attribute a of e_j.

In order to determine the affinity of a word and an event attribute, we create *attribute:value classifiers* – one for each *attribute:value pair* that occurs in any event. For example, for goals we create a *type:goal* classifier, and for unsuccessful events we create an *outcome:unsuccessful* classifier.

For the categorical attributes *Type, Outcome,* and *Pass Events*, we fit a linear SVM (Fan et al., 2008) using the utterance-level alignments provided by Ψ^{prior} (the exemplar SVMs) to determine the positive and negative examples. For instance, we use all the utterances which are aligned with an event whose *type* value is "pass" as positive examples for our *type:pass* classifier, and all other utterances as negative examples.

The weight assigned to each dimension in a linear SVM describes the relative importance of that dimension in the classification process. The dimensions of our *attribute:value* SVMs are the words of the corpus, normalized for case and minus punctuation and stop words. Therefore, the affinity of a word w_i and the *attribute:value* e_j^a is the weight of the dimension corresponding to w_i in the e_j^a *attribute:value* classifier. Following others (Liang et al., 2009; Kate and Mooney, 2007), we use string matches to determine the affinity between a word and the *Player* attribute.

In order to make comparisons between the importance of a word in the decision process for different classifiers, we normalize the weight vectors for each. These *attribute:value* classifiers produce our finest resolution alignments, allowing us to define a correspondence between a single word and a single attribute of any event.

By considering e_j in terms of its attributes, we are able to compose a score for e_j with fragment s. This is a kind of double-sided compositional semantics, where both the meaningful signs (s) and their extensions (e_j) are composed of finer-resolution atomic parts (w_i and e_j^a, respectively).

4.2 Segmentations Given Utterances

The potential function $\Phi^{\text{seg}}(s, W)$ from Equation 2 returns a score for a fragment within an utterance. A segmentation can be thought of as the collection of bigrams $\langle w_i, w_{i+1} \rangle$ where w_i is the last word of a fragment which is being used to describe one event and w_{i+1} is the first word of a fragment being used to describe a different event. We will refer to such bigrams as *splitpoints*.

The function Φ^{seg} should favor fragments that begin and end at good *splitpoint* and whose intermediate bigrams are bad *splitpoints*. We formalize this as follows:

$$\Phi^{\text{seg}}(s, W) \propto \frac{\phi(w_{k-1}, w_k) * \phi(w_{k+m}, w_{k+m+1})}{\prod_{j=0}^{m-1} \phi(w_{k+j}, w_{k+j+1})}$$

where fragment s is a span of m consecutive words $\{w_k, ..., w_{k+m}\}$ from W, and ϕ is a score for how good of a *splitpoint* $\langle w_i, w_{i+1} \rangle$ would make (explained below).

Ideally, ϕ will be a classifier which can tell us if a given bigram is a good *splitpoint* for the utterance W. However, ours being an attempt at weakly-supervised learning, we have no labeled examples of correct splitpoints from which

to work. Instead, we employ linguistic knowledge to create a proxy of labels. We will use this proxy to train a classifier to discover the features of good splitpoints which can be generalized and produce a more robust system.

The proxy labeling scheme we developed is based on conservative components common to a variety of theories of discourse. Discourse theories aim to model the relationships which exist between adjacent utterances in a coherent discourse. Since we consider a sports commentary to be a coherent discourse, we can leverage results from discourse theory in producing our proxy labels.

Temporal Discourse: Events in a soccer match occur in a temporal sequence, and so it is reasonable to assume that the language used to describe them will employ temporal discourse relations to distinguish fragments describing separate events. Pitler et al. (2008) have constructed a list of discourse relations which can be easily automatically identified, including temporal discourse relations. These are indicated by the presence of *discourse markers* — alternately known as cue phrases. We hypothesize that cue phrases can be used to identify splitpoints and use them in our proxy labeling scheme. This method is not restricted to temporally related discourse: some contingency, expansion, and comparison relations are also analyzed as "easily identifiable". As such, our segmentation process can also be used to ground language into a world state where these relations would hold.

Prosodic Discourse: We also make use of prosodic discourse cues. Pierrehumbert and Hirschberg (1990) claim that intonational phrases play an important role in discourse segmentation. Therefore, we hypothesize that the edges of intonational phrases are very likely to correspond with correct splitpoints. Viewing the commentary transcriptions as a noisy channel of the actual speech signal, we can identify the intonational phrase boundaries with the punctuation inserted in the transcription process. Chafe (1988) confirms that punctuation in written language has a strong correspondence with intonational phrase boundaries, and an assumption like ours has been successfully implemented in speech synthesis systems (Black and Lenzo, 2000). Thus, we include bigrams containing punctuation as splitpoints in our proxy labels.

Feature Description for *splitpoint* classifier
Is w_i/w_{i+1} a discourse marker?
Is w_i/w_{i+1} punctuation?
Is w_i/w_{i+1} a player name?
Part of speech of w_i/w_{i+1}
Is one of w_i/w_{i+1} a dependent of the other?
Are w_i and w_{i+1} dependents of the same governor?
Dependency relations that hold across splitpoint
Height of w_i/w_{i+1} in the dependency tree
Difference in height of w_i/w_{i+1} in dependency tree
$\psi(w_i, e_j)$ of all words left versus right of splitpoint
Symmetric difference of best affinity scores for w_i/w_{i+1}
Are best affinity scores from the same event?

Table 1: Feature description for splitpoint classifier

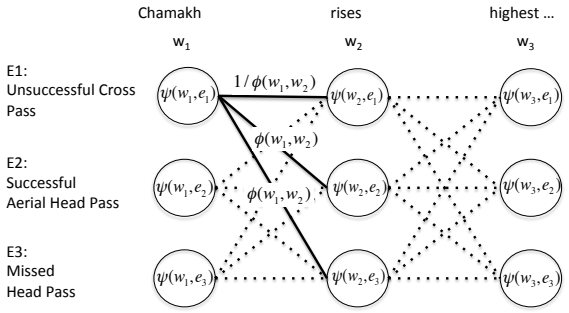

Figure 4: We use a trellis to allow for dynamic programming optimization of the objective function

Splitpoint Classifier: All other bigrams besides those above are labeled as negative examples, and a linear SVM is fit to the data. The features for the classifier include structural, discourse, and statistical features. We make use of dependency parse information from the Stanford dependency parser (De Marneffe and Manning, 2008). The full features list is explained in Table 1.

4.3 Optimization

We want to maximize the function in Equation 1, and we have explained that we can approximate this by maximizing the factored form in Equation 2. By the above methods, we can produce values for the functions Ψ^{align} and Φ^{seg}. What remains is to optimize Equation 2.

We take advantage of the factorization by using a dynamic programming approach to optimization. Figure 4 illustrates the setup. For each word w_i of the utterance, we create a column of nodes in our trellis, with one row for each event $e_j \in E$. The nodes represent the affinity of a given word w_i with event e_j. The weights on these nodes come from $\psi^{\text{atr.}}(w_i, e_j)$ described in section 4.2.

The nodes in column w_i are connected to the nodes in column w_{i+1} by edges whose weights

Method	Precision	Recall	F1
Liang et al. (2009)	0.513	0.393	0.445
Our approach	**0.603**	**0.481**	**0.535**

Table 2: Fragment-level alignments starting from gold utterance-level alignments

Method	Precision	Recall	F1
Liang et al. (2009)	0.211	0.135	0.165
Our approach	**0.235**	**0.255**	**0.245**

Table 3: Fragment-level alignments starting from raw data

are drawn from the splitpoint classifier response $\phi(w_i, w_{i+1})$. We label the edges between adjacent nodes corresponding to different events with the responses from the splitpoint classifier, and the inverse of these responses for edges connecting nodes corresponding to the same event.

We then use the Viterbi algorithm (Viterbi, 1967) to find the maximum scoring path through this trellis. The maximum scoring path optimizes Equation 2, and serves as our approximation of the optimization of Equation 1. We choose the top k diverse paths through the trellis and use the associations therein as our alignments. See Figure 5 for a detailed example of how our Viterbi path coincides with the responses from the *attribute:value* classifiers.

5 Experiments

One justification for multi-resolution language grounding would be if finer-resolution grounding improves coarser-resolution grounding and vice versa. If so, we expect that better utterance-level alignments will improve fragment-level alignments, and that in turn those fragment-level alignments will improve utterance-level alignments. We evaluate both of these hypotheses.

5.1 Experimental Setup

Dataset: We use the publicly available Professional Soccer Commentary (PSC) dataset introduced in Hajishirzi et al. (2012). This dataset is composed of professional commentaries from the 2010-2011 season of the English Premier League, along with a human-annotated data feed produced for each game by Opta Sportsdata (Opta, 2012) which describes all events occurring around the ball. Events include passes, shots, misses, cards, tackles, and other relevant game details. Each event category is defined precisely and the feed is annotated by professionals according to strict event description guidelines.

Method	Precision	Recall	F1
Liang et al. (2009)	0.327	0.418	0.367
Hajishirzi et al. (2012)	0.355	**0.576**	0.439
Our approach	**0.407**	0.520	**0.457**

Table 4: Utterance-level alignment results

The PSC also provides ground truth alignment of full utterances to events in the data feed, and for this work we have augmented it with ground truth fragment-level annotations[5].

We use data from 7 games of the PSC. These games consist of 778 utterances totaling 13,692 words. There are 12,275 events. This data is labeled with ground truth utterance- and fragment-alignments.

Metric: There are 1,295 correct utterance-to-event alignments. For evaluation we use precision, recall, and F1 of our utterance-level alignments.

The evaluation of fragment-level alignments is less straight forward. This is due to the two features of a correct fragment alignment: picking the correct fragment boundaries and associating the fragment with the correct event. We evaluate fragment-level alignment on a per word basis. We consider precision in this task to be the number of correct word to event alignments versus the total number of alignments produced by a system. Recall is the number of correct word to event alignments versus the total gold word to event alignments, of which there are 18,147.

Comparisons: We compare to two previous works: Liang et al. (2009), which produces both segmentation and alignment results; and Hajishirzi et al. (2012), which produces state-of-the-art alignments. When evaluating segmentation, we compare how well the systems perform starting from the raw dataset, and starting from gold utterance-level alignments. This allows us to isolate the segmentation process from the overall system architectures. It also gives us some insight into the effect of event priors on the segmentation and alignment processes.

[5]The full dataset is available at `http://ssli.ee.washington.edu/tial/projects/multires/`

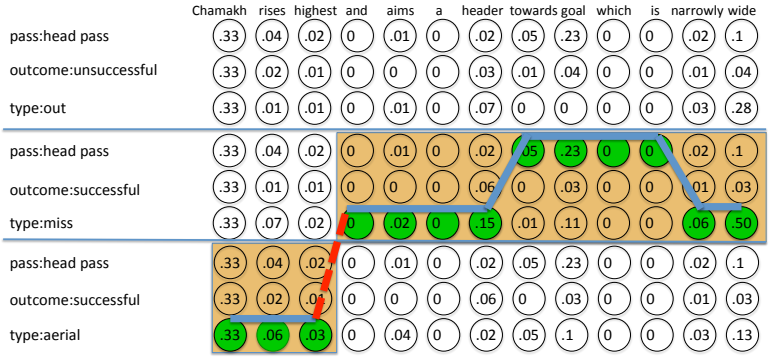

Figure 5: A successful grounding at multiple resolutions. Thin blue lines separate the *attribute:value* pairs corresponding to the three events. Values of $\psi(w_i, e_j)$ are shown on each node. The shaded bands indicate the gold fragment-level alignments. Thick line connecting the green nodes indicates the classifier responses used in the Viterbi best path through our trellis. The red dashed edge indicates a high response from the splitpoint classifier. This figure is best viewed in color.

Method	Precision	Recall	F1
Ours	0.235	0.255	0.245
- Ψ^{affinity}	0.213	0.133	0.164
- Φ^{seg}	0.205	0.232	0.218

Table 5: Ablation studies for fragment-level alignments by removing Ψ^{affinity} and Φ^{seg} from our model by replacing them with uniform function.

Method	Precision	Recall	F1
Ours	0.407	0.520	0.457
- Ψ^{affinity}	0.446	0.189	0.265
- Φ^{seg}	0.376	0.563	0.451

Table 6: Ablation studies for utterance-level alignments by removing Ψ^{affinity} and Φ^{seg} from our model by replacing them with uniform function.

5.2 Results

We evaluate our method on its alignments at the fragment-level and at the utterance-level. The results are as follows:

Fragment-level: Our results for segmentation can be seen in Tables 2 and 3. Table 2 shows the results achieved on the fragment-level alignment task using human-labeled utterance to event alignments. In this setting, all and only the correct events for each utterance are present. Still, there are several ambiguities in the data. Some fragments are aligned in the gold data with multiple events, and some are aligned to no event. Our method outperforms the previous by a large margin in terms of both precision and recall. We show below how this is due to our system's accommodation of discourse structure when making segmentation decisions and the factored form of our optimization.

Table 3 shows the results for fragment-level alignment by applying each system starting from the raw data. Here, in addition to the ambiguities mentioned above, the problem is further complicated by the fact that some correct events are missing from the alignments produced by each system and some incorrect events are included in these alignments (see Error Analysis below for details). Still our method achieves a significant improvement, with a 48% increase in F1 versus prior work.

Table 5 shows ablation results for the effect of the factors used in our optimization for fragment-level alignments. These results demonstrate the value of each factor in the fragment-level alignment process. We cannot ascribe the benefit of this method to one factor or another alone – it is their concert that improves performance.

Utterance-level: We have posited that good finer-resolution alignments will improve the coarser-resolution utterance to event alignments. Our results confirm this hypothesis. Table 4 shows our results on these alignments. We are able to improve F1 versus a state-of-the-art system which is tuned to maximize its F1 score. The majority of our improvement comes from the increased precision of our system, due to the influence of the finer-resolution fragment-level alignments on these coarser, utterance-level alignments. We provide a detailed example of this below. Ablation results are shown in Table 6.

5.3 Qualitative Analysis

A qualitative analysis of our system reveals the power of our factored objective, double-sided compositional approach, and leveraging of discourse structure. Figure 5 shows the best path through the trellis of the example sentence used in the introduction. For explanatory purposes, we have split every event into its three component attributes. This allows us to see how the *attribute:value* classifiers combine to produce an alignment.

Discourse Structure: The fragment-level alignment we have produced for this utterance is perfect: it correctly identifies the single *splitpoint* and correctly identifies each fragment with the associated event.

The identification of the splitpoint "and" comes from the fact that this word has, among other uses, a discourse connective meaning. Thus, the edges

in our trellis between different events are weighted higher than edges between the same event in the edges between the nodes for "highest" and "and", encouraging the Viterbi path to change events at this point.

Compositionality: We can see effect of the compositional approach we have taken – composing $\psi^{\text{affinity}}(s, e_j)$ from the *attribute:value* classifier scores of each $\psi(w_i, e_j^a)$ – by looking at how the best path makes use of different attributes of the same event. For the "miss" event aligned with the second part of the sentence, we can see that the best path makes use of both values from the *type:miss* and *pass event:head pass* classifiers.

Affinities: A few interesting associations are worth pointing out. First, we note that the word "header" has a stronger affinity for the *type:miss* attribute than it does for the *pass events:head pass* attribute. On first blush, this seems like a mistake in our classifier. However, we can see that even in this single trellis all three events have the *pass events:head pass* attribute. The utterance-level alignment uses this association already, aligning utterances containing the word "header" with events that have a *pass events:head pass* attribute. At a finer-resolution, it is necessary to make a different distinction between events. Our method finds that the presence of the word "header" is a stronger indicator of an event with a *type:miss* attribute, and thus this association is made.

Words that are better for the coarser-resolution association with the *pass events:head pass* attribute are "towards" and "goal". Out of the 10 utterances containing the word "towards" in the dataset, 3 of these are aligned with at least 1 *pass events:head pass* event, making this strong association a correct one. The word "goal" also has an affinity for the *pass events:head pass* attribute due to the fact that many events with this attribute are attempts on goal. This correlates with domain knowledge about soccer, because, although there may be other uses of their head by a player in the game, shots on goal are events which will nearly always be commented upon by an announcer.

Factorization: We have shown that finer-resolution fragment-level alignments can improve utterance-level alignments. From the exemplar SVMs, we are given an utterance-level alignment of the three events shown in the trellis with the utterance. This alignment is incorrect: the gold utterance alignment only includes the bottom two events. But by building an utterance-level alignment from the results of our fragment level alignment, we are left with only the two correct events. We prune the topmost event due to its failure to participate in a finer-resolution alignment.

5.4 Error Analysis

The majority of the errors made on our fragment-level alignments come in one of two flavors: Firstly, we sometimes erroneously identify a fragment as referring to an event when in truth it refers to no event. Commentators often describe facts about players or the weather or previous games which have no extension in the current game. However, our system cannot distinguish such language from the language referring to this game. This is a good avenue for future exploration.

The second set of errors we make in fragmentation are caused by bad event priors. Our current setup cannot increase recall: we can only improve the precision of the utterance-level alignments we are given. Therefore, if an event is overlooked in the first-pass of utterance-level alignments, we cannot reintroduce it through a fragment alignment. This is a direction for future work as well.

6 Related Work

Early semantic parsing work made use of fully supervised training (Zettlemoyer and Collins, 2005; Ge and Mooney, 2006; Snyder and Barzilay, 2007), but more recent work has focused on reducing the amount of supervision required (Artzi and Zettlemoyer, 2013). A few unsupervised approaches exist (Poon and Domingos, 2009; Poon, 2013), but these are specific to translating language into queries in highly structured database and cannot be applied to our more flexible domain.

There are few datasets as detailed as the Professional Soccer Commentary Dataset. Early work in understanding soccer commentaries focused on RoboCup soccer (Chen and Mooney, 2008; Chen et al., 2010; Bordes et al., 2010; Hajishirzi et al., 2011) where simple language describes each event, and events are in a one-to-one correspondence with utterances. Another dataset used for language grounding is the Weather Report Dataset (Liang et al., 2009). Here, again, however, we have mostly single utterances paired with single events, and many alignments are made via numerical string matching rather than learning lex-

ical cues. The NFL Recap dataset (Snyder and Barzilay, 2007) is also laden with numerical fact matching, and does not include the fragment-level segmentation annotation that the PSC dataset provides.

Impressive advances have been made grounding language in instructions. Branavan et al. (2009) and Vogel and Jurafsky (2010) work in the domain of computer technical support instructions, mapping language to actions using reinforcement learning. Matuszek et al. (2012b) parses simple language to robot control instructions. Our work focuses on dealing with a richer space, both in terms of the language used and the world-representation into which it is grounded, and leveraging the multiple resolutions of reference.

An exciting direction of research, closer to our own, aims to ground natural language in visual perception systems. Matuszek et al. (2012a) attempts to learn a joint model of language and object characteristics of a workplace environment. Yu and Siskind (2013) grounds moderately rich language in automatically annotated video clips. Again, the contribution of our work versus the above is in the complexity of the language with which we deal and our multi-resolution model.

7 Conclusion

The problem of grounding complex natural human language such as soccer commentaries is extremely difficult at all resolutions, and it is most challenging at finer resolutions where data is sparsest and small errors cannot be as easily normalized. Our work will help open new avenues of research into this difficult and exciting problem.

This paper presents a new method for the multi-resolution grounding of complex natural language in a detailed world representation. Our factor graph allows us to decompose the grounding problem into the more tractable subproblems of segmenting the language into fragments and aligning the fragments with the world representation. In the segmentation phase, we make use of linguistic theories of discourse to create a proxy of labels from which we learn statistical and structural features of good splitpoints. In the alignment phase, we bootstrap the learning of finer-grained correspondences between the language and the world representation with rough alignments from a state-of-the-art system. We combine these phases in a dynamic programming setup which allows us to efficiently optimize our objective.

We have shown that factoring the acquisition problem into separate alignment and segmentation phases improves performance on several evaluation metrics. We achieve considerable improvements over the previous state of the art on finer-resolution alignments in the domain of professional soccer commentaries, and we show that we can leverage groundings at one resolution to improve alignments in another.

Several extensions of this work are possible. We would like to annotate more games to improve our dataset. We could improve our model by encoding the dynamics of the environment. We did not attempt to learn this information in our process, but it is likely that modeling the event transition probabilities could provide better results. A larger future work would extend the method outlined herein to produce templates for automated commentary generation.

Acknowledgments

This research was supported in part by a grant from the NSF (IIS-1352249), and the Royalty Research Fund (RRF) at the University of Washington. The authors also wish to thank Gina-Anne Levow, Yoav Artzi, Ben Hixon, and the anonymous reviewers for their valuable feedback on this work.

References

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1(1):49–62.

Alan Black and Kevin Lenzo. 2000. Building voices in the festival speech synthesis system.

Antoine Bordes, Nicolas Usunier, and Jason Weston. 2010. Label ranking under ambiguous supervision for learning semantic correspondences. In *Proceedings of The 27th International Conference on Machine Learning*, pages 103–110.

S. R. K. Branavan, Harr Chen, Luke Zettlemoyer, and Regina Barzilay. 2009. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 82–90.

Wallace Chafe. 1988. Punctuation and the prosody of written language. *Written communication*, 5(4):395–426.

David L. Chen and Raymond J. Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th International Conference on Machine Learning*, pages 128–135.

David L. Chen, Joohyun Kim, and Raymond J. Mooney. 2010. Training a multilingual sportscaster: Using perceptual context to learn language. *Journal of Artificial Intelligence Research (JAIR)*, 37:397–435.

Donald Davidson. 1967. The logical form of action sentences. *The logic of Decision and Action*.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. Stanford typed dependencies manual. *URL http://nlp. stanford. edu/software/dependencies manual. pdf*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Ruifang Ge and Raymond J. Mooney. 2006. Discriminative reranking for semantic parsing. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*.

Hannaneh Hajishirzi, Julia Hockenmaier, Erik T. Mueller, and Eyal Amir. 2011. Reasoning about robocup soccer narratives. In *Proceedings of the 27th conference on Uncertainty in Artificial Intelligence*, pages 291–300.

Hannaneh Hajishirzi, Mohammad Rastegari, Ali Farhadi, and Jessica K Hodgins. 2012. Semantic understanding of professional soccer commentaries. In *Proceedings of the 28th conference on Uncertainty in Artificial Intelligence*.

Rohit J. Kate and Raymond J. Mooney. 2007. Learning language semantics from ambiguous supervision. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, pages 895–900.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 91–99.

Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. 2011. Ensemble of exemplar-svms for object detection and beyond. In *Proceedings of the 13th International Conference on Computer Vision.*

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012a. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proc. of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland, June.

Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. 2012b. Learning to parse natural language commands to a robot control system. In *Proc. of the 13th International Symposium on Experimental Robotics (ISER)*, June.

Opta. 2012. http://www.optasports.com.

Janet Pierrehumbert and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. *Intentions in Communication*, 271.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K Joshi. 2008. Easily identifiable discourse relations. *Technical Reports (CIS)*, page 884.

Hoifung Poon and Pedro Domingos. 2009. Unsupervised semantic parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1–10.

Hoifung Poon. 2013. Grounded unsupervised semantic parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. 2011. Data-driven visual similarity for cross-domain image matching. *ACM Transaction of Graphics (TOG) (Proceedings of ACM SIGGRAPH ASIA)*, 30(6).

Benjamin Snyder and Regina Barzilay. 2007. Database-text alignment via structured multilabel classification. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1713–1718.

Andrew J Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269.

Adam Vogel and Daniel Jurafsky. 2010. Learning to follow navigational directions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 806–814.

Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 53–63.

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, pages 658–666.