Bootstrapping Semantic Analyzers from Non-Contradictory Texts

Ivan Titov **Mikhail Kozhevnikov**
Saarland University
Saarbrücken, Germany
{titov|m.kozhevnikov}@mmci.uni-saarland.de

Abstract

We argue that groups of unannotated texts with overlapping and non-contradictory semantics represent a valuable source of information for learning semantic representations. A simple and efficient inference method recursively induces joint semantic representations for each group and discovers correspondence between lexical entries and latent semantic concepts. We consider the generative semantics-text correspondence model (Liang et al., 2009) and demonstrate that exploiting the non-contradiction relation between texts leads to substantial improvements over natural baselines on a problem of analyzing human-written weather forecasts.

1 Introduction

In recent years, there has been increasing interest in statistical approaches to semantic parsing. However, most of this research has focused on supervised methods requiring large amounts of labeled data. The supervision was either given in the form of meaning representations aligned with sentences (Zettlemoyer and Collins, 2005; Ge and Mooney, 2005; Mooney, 2007) or in a somewhat more relaxed form, such as lists of candidate meanings for each sentence (Kate and Mooney, 2007; Chen and Mooney, 2008) or formal representations of the described world state for each text (Liang et al., 2009). Such annotated resources are scarce and expensive to create, motivating the need for unsupervised or semi-supervised techniques (Poon and Domingos, 2009). However, unsupervised methods have their own challenges: they are not always able to discover semantic equivalences of lexical entries or logical forms or, on the contrary, cluster semantically different or even opposite expressions (Poon and Domingos, 2009). Unsupervised approaches can only rely on distributional similarity of contexts (Harris, 1968) to decide on semantic relatedness of terms, but this information may be sparse and not reliable (Weeds and Weir, 2005). For example, when analyzing weather forecasts it is very hard to discover in an unsupervised way which of the expressions among "south wind", "wind from west" and "southerly" denote the same wind direction and which are not, as they all have a very similar distribution of their contexts. The same challenges affect the problem of identification of argument roles and predicates.

In this paper, we show that groups of unannotated texts with overlapping and non-contradictory semantics provide a valuable source of information. This form of weak supervision helps to discover implicit clustering of lexical entries and predicates, which presents a challenge for purely unsupervised techniques. We assume that each text in a group is independently generated from a full latent semantic state corresponding to the group. Importantly, the texts in each group do not have to be paraphrases of each other, as they can verbalize only specific parts (*aspects*) of the full semantic state, yet statements about the same aspects must not contradict each other. Simultaneous inference of the semantic state for the non-contradictory and semantically overlapping documents would restrict the space of compatible hypotheses, and, intuitively, 'easier' texts in a group will help to analyze the 'harder' ones.[1]

As an illustration of why this weak supervision may be valuable, consider a group of two non-contradictory texts, where one text mentions "2.2 bn GBP decrease in profit", whereas another one includes a passage "profit fell by 2.2 billion pounds". Even if the model has not observed

[1] This view on this form of supervision is evocative of co-training (Blum and Mitchell, 1998) which, roughly, exploits the fact that the same example can be 'easy' for one model but 'hard' for another one.

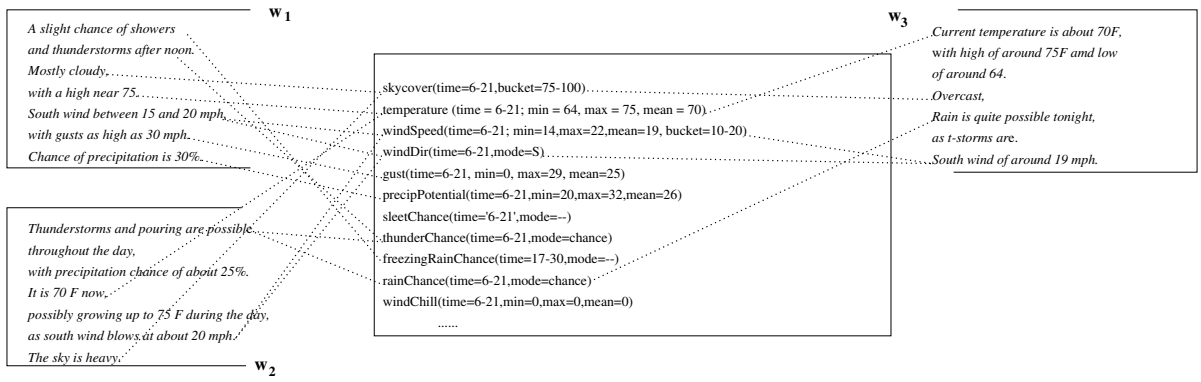

Figure 1: An example of three non-contradictory weather forecasts and their alignment to the semantic representation. Note that the semantic representation (the block in the middle) is not observable in training.

the word "fell" before, it is likely to align these phrases to the same semantic form because of similarity of their arguments. And this alignment would suggest that "fell" and "decrease" refer to the same process, and should be clustered together. This would not happen for the pair "fell" and "increase" as similarity of their arguments would normally entail contradiction. Similarly, in the example mentioned earlier, when describing a forecast for a day with expected south winds, texts in the group can use either "south wind" or "southerly" to indicate this fact but no texts would verbalize it as "wind from west", and therefore these expressions will be assigned to different semantic clusters. However, it is important to note that the phrase "wind from west" may still appear in the texts, but in reference to other time periods, underlying the need for modeling alignment between grouped texts and their latent meaning representation.

As much of the human knowledge is redescribed multiple times, we believe that non-contradictory and semantically overlapping texts are often easy to obtain. For example, consider semantic analysis of news articles or biographies. In both cases we can find groups of documents referring to the same events or persons, and though they will probably focus on different aspects and have different subjective passages, they are likely to agree on the core information (Shinyama and Sekine, 2003). Alternatively, if such groupings are not available, it may still be easier to give each semantic representation (or a state) to multiple annotators and ask each of them to provide a textual description, instead of annotating texts with semantic expressions. The state can be communicated to them in a visual or audio form (e.g., as a picture or a short video clip) ensuring that their interpretations are consistent.

Unsupervised learning with shared latent semantic representations presents its own challenges, as exact inference requires marginalization over possible assignments of the latent semantic state, consequently, introducing non-local statistical dependencies between the decisions about the semantic structure of each text. We propose a simple and fairly general approximate inference algorithm for probabilistic models of semantics which is efficient for the considered model, and achieves favorable results in our experiments.

In this paper, we do not consider models which aim to produce complete formal meaning of text (Zettlemoyer and Collins, 2005; Mooney, 2007; Poon and Domingos, 2009), instead focusing on a simpler problem studied in (Liang et al., 2009). They investigate grounded language acquisition set-up and assume that semantics (*world state*) can be represented as a set of *records* each consisting of a set of *fields*. Their model segments text into utterances and identifies records, fields and field values discussed in each utterance. Therefore, one can think of this problem as an extension of the semantic role labeling problem (Carreras and Marquez, 2005), where predicates (i.e. *records* in our notation) and their arguments should be identified in text, but here arguments are not only assigned to a specific role (*field*) but also mapped to an underlying equivalence class (*field value*). For example, in the weather forecast domain field *sky cover* should get the same value given expressions "overcast" and "very cloudy" but a different one if the expres-

sions are "clear" or "sunny". This model is hard to evaluate directly as text does not provide information about all the fields and does not necessarily provide it at the sufficient granularity level. Therefore, it is natural to evaluate their model on the database-text alignment problem (Snyder and Barzilay, 2007), i.e. measuring how well the model predicts the alignment between the text and the observable records describing the entire world state. We follow their set-up, but assume that instead of having access to the full semantic state for every training example, we have a very small amount of data annotated with semantic states and a larger number of unannotated texts with non-contradictory semantics.

We study our set-up on the weather forecast data (Liang et al., 2009) where the original textual weather forecasts were complemented by additional forecasts describing the same weather states (see figure 1 for an example). The average overlap between the verbalized fields in each group of non-contradictory forecasts was below 35%, and more than 60% of fields are mentioned only in a single forecast from a group. Our model, learned from 100 labeled forecasts and 259 groups of unannotated non-contradictory forecasts (750 texts in total), achieved 73.9% F_1. This compares favorably with 69.1% shown by a semi-supervised learning approach, though, as expected, does not reach the score of the model which, in training, observed semantics states for all the 750 documents (77.7% F_1).

The rest of the paper is structured as follows. In section 2 we describe our inference algorithm for groups of non-contradictory documents. Section 3 redescribes the semantics-text correspondence model (Liang et al., 2009) in the context of our learning scenario. In section 4 we provide an empirical evaluation of the proposed method. We conclude in section 5 with an examination of additional related work.

2 Inference with Non-Contradictory Documents

In this section we will describe our inference method on a higher conceptual level, not specifying the underlying meaning representation and the probabilistic model. An instantiation of the algorithm for the semantics-text correspondence model is given in section 3.2.

Statistical models of parsing can often be regarded as defining the probability distribution of meaning $\boldsymbol{m}$ and its alignment $\boldsymbol{a}$ with the given text $\boldsymbol{w}$, $P(\boldsymbol{m}, \boldsymbol{a}, \boldsymbol{w}) = P(\boldsymbol{a}, \boldsymbol{w}|\boldsymbol{m})P(\boldsymbol{m})$. The semantics $\boldsymbol{m}$ can be represented either as a logical formula (see, e.g., (Poon and Domingos, 2009)) or as a set of field values if database records are used as a meaning representation (Liang et al., 2009). The alignment $\boldsymbol{a}$ defines how semantics is verbalized in the text $\boldsymbol{w}$, and it can be represented by a meaning derivation tree in case of full semantic parsing (Poon and Domingos, 2009) or, e.g., by a hierarchical segmentation into utterances along with an utterance-field alignment in a more shallow variation of the problem. In semantic parsing, we aim to find the most likely underlying semantics and alignment given the text:

$$(\hat{\boldsymbol{m}}, \hat{\boldsymbol{a}}) = \arg\max_{\boldsymbol{m},\boldsymbol{a}} P(\boldsymbol{a}, \boldsymbol{w}|\boldsymbol{m})P(\boldsymbol{m}). \quad (1)$$

In the supervised case, where $\boldsymbol{a}$ and $\boldsymbol{m}$ are observable, estimation of the generative model parameters is generally straightforward. However, in a semi-supervised or unsupervised case variational techniques, such as the EM algorithm (Dempster et al., 1977), are often used to estimate the model. As common for complex generative models, the most challenging part is the computation of the posterior distributions $P(\boldsymbol{a}, \boldsymbol{m}|\boldsymbol{w})$ on the E-step which, depending on the underlying model $P(\boldsymbol{m}, \boldsymbol{a}, \boldsymbol{w})$, may require approximate inference.

As discussed in the introduction, our goal is to integrate groups of non-contradictory documents into the learning procedure. Let us denote by $\boldsymbol{w}_1,\ldots,\boldsymbol{w}_K$ a group of non-contradictory documents. As before, the estimation of the posterior probabilities $P(\boldsymbol{m}_i, \boldsymbol{a}_i|\boldsymbol{w}_1 \ldots \boldsymbol{w}_K)$ presents the main challenge. Note that the decision about $\boldsymbol{m}_i$ is now conditioned on all the texts $\boldsymbol{w}_j$ rather than only on $\boldsymbol{w}_i$. This conditioning is exactly what drives learning, as the information about likely semantics $\boldsymbol{m}_j$ of text j affects the decision about choice of $\boldsymbol{m}_i$:

$$P(\boldsymbol{m}_i|\boldsymbol{w}_1,\ldots,\boldsymbol{w}_K) \propto \sum_{a_i} P(\boldsymbol{a}_i, \boldsymbol{w}_i|\boldsymbol{m}_i) \times$$
$$\times \sum_{\boldsymbol{m}_{-i},\boldsymbol{a}_{-i}} P(\boldsymbol{m}_i|\boldsymbol{m}_{-i})P(\boldsymbol{m}_{-i}, \boldsymbol{a}_{-i}, \boldsymbol{w}_{-i}), \quad (2)$$

where $\boldsymbol{x}_{-i}$ denotes $\{\boldsymbol{x}_j : j \neq i\}$. $P(\boldsymbol{m}_i|\boldsymbol{m}_{-i})$ is the probability of the semantics $\boldsymbol{m}_i$ given all the meanings $\boldsymbol{m}_{-i}$. This probability assigns zero weight to inconsistent meanings, i.e. such mean-

ings $(\boldsymbol{m_1},...,\boldsymbol{m_K})$ that $\wedge_{i=1}^{K}\boldsymbol{m}_i$ is not satisfiable,[2] and models dependencies between components in the composite meaning representation (e.g., arguments values of predicates). As an illustration, in the forecast domain it may express that clouds, and not sunshine, are likely when it is raining. Note, that this probability is different from the probability that $\boldsymbol{m}_i$ is actually verbalized in the text.

Unfortunately, these dependencies between $\boldsymbol{m}_i$ and $\boldsymbol{w}_j$ are non-local. Even though the dependencies are only conveyed via $\{\boldsymbol{m}_j : j \neq i\}$ the space of possible meanings $\boldsymbol{m}$ is very large even for relatively simple semantic representations, and, therefore, we need to resort to efficient approximations.

One natural approach would be to use a form of belief propagation (Pearl, 1982; Murphy et al., 1999), where messages pass information about likely semantics between the texts. However, this approach is still expensive even for simple models, both because of the need to represent distributions over $\boldsymbol{m}$ and also because of the large number of iterations of message exchange needed to reach convergence (if it converges).

An even simpler technique would be to parse texts in a random order conditioning each meaning $\boldsymbol{m}_k^\star$ for $k \in \{1,...,K\}$ on all the previous semantics $\boldsymbol{m}_{<k}^\star = \boldsymbol{m}_1^\star,...,\boldsymbol{m}_{k-1}^\star$:

$$\boldsymbol{m}_k^\star = \arg\max_{\boldsymbol{m}_k} P(\boldsymbol{w}_k|\boldsymbol{m}_k)P(\boldsymbol{m}_k|\boldsymbol{m}_{<k}^\star).$$

Here, and in further discussion, we assume that the above search problem can be efficiently solved, exactly or approximately. However, a major weakness of this algorithm is that decisions about components of the composite semantic representation (e.g., argument values) are made only on the basis of a single text, which first mentions the corresponding aspects, without consulting any future texts $k' > k$, and these decisions cannot be revised later.

We propose a simple algorithm which aims to find an appropriate order of the greedy inference by estimating how well each candidate semantics $\hat{\boldsymbol{m}}_k$ would explain other texts and at each step selecting k (and $\hat{\boldsymbol{m}}_k$) which explains them best.

The algorithm, presented in figure 2[3], constructs an ordering of texts $\boldsymbol{n} = (n_1,...,n_K)$

1: $\boldsymbol{n} := (), \boldsymbol{m}^\star := ()$
2: **for** $i := 1 : K-1$ **do**
3: **for** $j \notin \boldsymbol{n}$ **do**
4: $\hat{\boldsymbol{m}}_j := \arg\max_{\boldsymbol{m}_j} P(\boldsymbol{m}_j, \boldsymbol{w}_j|\boldsymbol{m}^\star)$
5: **end for**
6: $n_i := \arg\max_{j\notin\boldsymbol{n}} P(\hat{\boldsymbol{m}}_j, \boldsymbol{w}_j|\boldsymbol{m}^\star)\times$
$\times \prod_{k\notin\boldsymbol{n}\cup\{j\}} \max_{\boldsymbol{m}_k} P(\boldsymbol{m}_k, \boldsymbol{w}_k|\boldsymbol{m}^\star, \hat{\boldsymbol{m}}_j)$
7: $\boldsymbol{m}_i^\star := \hat{\boldsymbol{m}}_{n_i}$
8: **end for**
9: $n_K := \{1,...,K\}\backslash\boldsymbol{n}$
10: $\boldsymbol{m}_K^\star := \arg\max_{\boldsymbol{m}_{n_K}} P(\boldsymbol{m}_{n_K}, \boldsymbol{w}_{n_K}|\boldsymbol{m}^\star)$

Figure 2: The approximate inference algorithm.

and corresponding meaning representations $\boldsymbol{m}^\star = (\boldsymbol{m}_1^\star,...,\boldsymbol{m}_K^\star)$, where $\boldsymbol{m}_k^\star$ is the predicted meaning representation of text $\boldsymbol{w}_{n_k}$. It starts with an empty ordering $\boldsymbol{n} = ()$ and an empty list of meanings $\boldsymbol{m}^\star = ()$ (line 1). Then it iteratively predicts meaning representations $\hat{\boldsymbol{m}}_j$ conditioned on the list of semantics $\boldsymbol{m}^\star = (\boldsymbol{m}_1^\star,...,\boldsymbol{m}_{i-1}^\star)$ fixed on the previous stages and does it for all the remaining texts $\boldsymbol{w}_j$ (lines 3-5). The algorithm selects a single meaning $\hat{\boldsymbol{m}}_j$ which maximizes the probability of all the remaining texts and excludes the text j from future consideration (lines 6-7).

Though the semantics $\boldsymbol{m}_k$ ($k \notin \boldsymbol{n}\cup\{j\}$) used in the estimates (line 6) can be inconsistent with each other, the final list of meanings $\boldsymbol{m}^\star$ is guaranteed to be consistent. It holds because on each iteration we add a single meaning $\hat{\boldsymbol{m}}_{n_i}$ to $\boldsymbol{m}^\star$ (line 7), and $\hat{\boldsymbol{m}}_{n_i}$ is guaranteed to be consistent with $\boldsymbol{m}^\star$, as the semantics $\hat{\boldsymbol{m}}_{n_i}$ was conditioned on the meaning $\boldsymbol{m}^\star$ during inference (line 4).

An important aspect of this algorithm is that unlike usual greedy inference, the remaining ('future') texts do affect the choice of meaning representations made on the earlier stages. As soon as semantics $\boldsymbol{m}_k^\star$ are inferred for every k, we find ourselves in the set-up of learning with unaligned semantic states considered in (Liang et al., 2009).

The induced alignments $\boldsymbol{a}_1,...,\boldsymbol{a}_K$ of semantics $\boldsymbol{m}^\star$ to texts $\boldsymbol{w}_1,...,\boldsymbol{w}_K$ at the same time induce alignments between the texts. The problem of producing multiple sequence alignment, especially in the context of sentence alignments, has been extensively studied in NLP (Barzilay and Lee, 2003). In this paper, we use semantic structures as a pivot for finding the best alignment in the hope that presence of meaningful text alignments will improve the quality of the resulting semantic structures by enforcing a form of agreement between them.

[2]Note that checking for satisfiability may be expensive or intractable depending on the formalism.

[3]We slightly abuse notation by using set operations with the lists $\boldsymbol{n}$ and $\boldsymbol{m}^\star$ as arguments. Also, for all the document indices j we use $j \notin S$ to denote $j \in \{1,...,K\}\backslash S$.

3 A Model of Semantics

In this section we redescribe the semantics-text correspondence model (Liang et al., 2009) with an extension needed to model examples with latent states, and also explain how the inference algorithm defined in section 2 can be applied to this model.

3.1 Model definition

Liang et al. (2009) considered a scenario where each text was annotated with a world state, even though alignment between the text and the state was not observable. This is a weaker form of supervision than the one traditionally considered in supervised semantic parsing, where the alignment is also usually provided in training (Chen and Mooney, 2008; Zettlemoyer and Collins, 2005). Nevertheless, both in training and testing the world state is observable, and the alignment and the text are conditioned on the state during inference. Consequently, there was no need to model the distribution of the world state. This is different for us, and we augment the generative story by adding a simplistic world state generation step.

As explained in the introduction, the world states $\boldsymbol{s}$ are represented by sets of records (see the block in the middle of figure 1 for an example of a world state). Each record is characterized by a record type $t \in \{1,...,T\}$, which defines the set of fields $\boldsymbol{F}^{(t)}$. There are $n^{(t)}$ records of type t and this number may change from document to document. For example, there may be more than a single record of type *wind speed*, as they may refer to different time periods but all these records have the same set of fields, such as minimal, maximal and average wind speeds. Each field has an associated type: in our experiments we consider only categorical and integer fields. We write $s^{(t)}_{n,f} = v$ to denote that n-th record of type t has field f set to value v.

Each document k verbalizes a subset of the entire world state, and therefore semantics $\boldsymbol{m}_k$ of the document is an assignment to $|\boldsymbol{m}_k|$ verbalized fields: $\wedge_{q=1}^{|\boldsymbol{m}_k|}(s^{(t_q)}_{n_q,f_q} = v_q)$, where t_q, n_q, f_q are the verbalized record types, records and fields, respectively, and v_q is the assigned field value. The probability of meaning m_k then equals the probability of this assignment with other state variables left non-observable (and therefore marginalized out). In this formalism checking for contradiction is trivial: two meaning representations contradict each other if they assign different values to the same field of the same record.

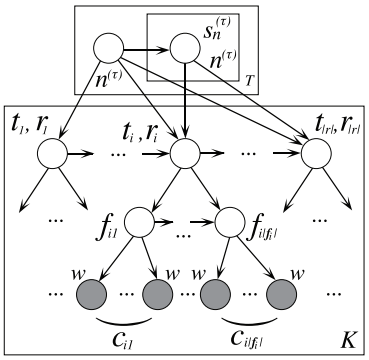

Figure 3: The semantics-text correspondence model with K documents sharing the same latent semantic state.

The semantics-text correspondence model defines a hierarchical segmentation of text: first, it segments the text into fragments discussing different records, then the utterances corresponding to each record are further segmented into fragments verbalizing specific fields of that record. An example of a segmented fragment is presented in figure 4. The model has a designated null-record which is aligned to words not assigned to any record. Additionally there is a null-field in each record to handle words not specific to any field. In figure 3 the corresponding graphical model is presented. The formal definition of the model for documents $\boldsymbol{w_1},..., \boldsymbol{w_K}$ sharing a semantic state is as follows:

- Generation of world state $\boldsymbol{s}$:
  - For each type $\tau \in \{1,...,T\}$ choose a number of records of that type $n^{(\tau)} \sim \text{Unif}(1,..., n_{max})$.
  - For each record $\boldsymbol{s}^{(\tau)}_n$, $n \in \{1,..,n^{(\tau)}\}$ choose field values $s^{(\tau)}_{nf}$ for all fields $f \in \boldsymbol{F}^{(\tau)}$ from the type-specific distribution.
- Generation of the verbalizations, for each document $\boldsymbol{w}_k$, $k \in \{1,..., K\}$:[4]
  - Record Types: Choose a sequence of verbalized record types $\boldsymbol{t} = (t_1,..., t_{|\boldsymbol{t}|})$ from the first-order Markov chain.
  - Records: For each type t_i choose a verbalized record $\boldsymbol{r}_i$ from all the records of that type: $l \sim \text{Unif}(1,..., n^{(\tau)})$, $\boldsymbol{r}_i := \boldsymbol{s}^{(t_i)}_l$.
  - Fields: For each record $\boldsymbol{r}_i$ choose a sequence of verbalized fields $\boldsymbol{f}_i = (f_{i1},..., f_{i|\boldsymbol{f}_i|})$ from the first-order Markov chain ($f_{ij} \in \boldsymbol{F}^{(t_i)}$).
  - Length: For each field f_{ij}, choose length $c_{ij} \sim \text{Unif}(1,..., c_{max})$.
  - Words: Independently generate c_{ij} words from the field-specific distribution $P(w|f_{ij}, r_{if_{ij}})$.

[4] We omit index k in the generative story and figure 3 to simplify the notation.

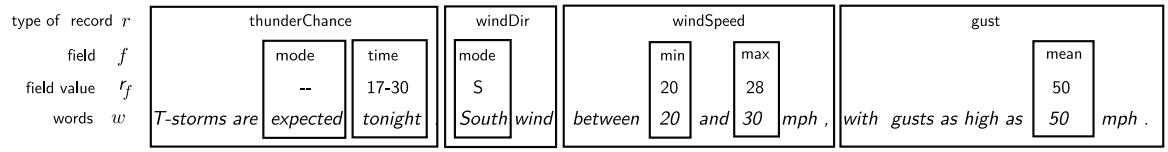

Figure 4: A segmentation of a text fragment into records and fields.

Note that, when generating fields, the Markov chain is defined over fields and the transition parameters are independent of the field values $r_{if_{ij}}$. On the contrary, when drawing a word, the distribution of words is conditioned on the value of the corresponding field.

The form of word generation distributions $P(w|f_{ij}, r_{if_{ij}})$ depends on the type of the field $f_{i,j}$. For categorical fields, the distribution of words is modeled as a distinct multinomial for each field value. Verbalizations of numerical fields are generated via a perturbation on the field value $r_{if_{ij}}$: the value $r_{if_{ij}}$ can be perturbed by either rounding it (up or down) or distorting (up or down, modeled by a geometric distribution). The parameters corresponding to each form of generation are estimated during learning. For details on these emission models, as well as for details on modeling record and field transitions, we refer the reader to the original publication (Liang et al., 2009).

In our experiments, when choosing a world state $\boldsymbol{s}$, we generate the field values independently. This is clearly a suboptimal regime as often there are very strong dependencies between field values: e.g., in the weather domain many record types contain groups of related fields defining minimal, maximal and average values of some parameter. Extending the method to model, e.g., pairwise dependencies between field values is relatively straightforward.

As explained above, semantics of a text $\boldsymbol{m}$ is defined by the assignment of state variables $\boldsymbol{s}$. Analogously, an alignment $\boldsymbol{a}$ between semantics $\boldsymbol{m}$ and a text $\boldsymbol{w}$ is represented by all the remaining latent variables: by the sequence of record types $\boldsymbol{t} = (t_1,...,t_{|\boldsymbol{t}|})$, choice of records $\boldsymbol{r}_i$ for each t_i, the field sequence $\boldsymbol{f}_i$ and the segment length c_{ij} for every field f_{ij}.

3.2 Learning and inference

We select the model parameters θ by maximizing the marginal likelihood of the data, where the data $\mathcal{D}$ is given in the form of groups $\boldsymbol{w} = \{\boldsymbol{w}_1,...,\boldsymbol{w}_K\}$ sharing the same latent state:[5]

$$\max_{\theta} \prod_{\boldsymbol{w}\in\mathcal{D}} \sum_{\boldsymbol{s}} P(\boldsymbol{s}) \prod_{k} \sum_{\boldsymbol{r},\boldsymbol{f},\boldsymbol{c}} P(\boldsymbol{r},\boldsymbol{f},\boldsymbol{c},\boldsymbol{w}_k|\boldsymbol{s},\theta).$$

To estimate the parameters, we use the Expectation-Maximization algorithm (Dempster et al., 1977). When the world state is observable, learning does not require any approximations, as dynamic programming (a form of the forward-backward algorithm) can be used to infer the posterior distribution on the E-step (Liang et al., 2009). However, when the state is latent, dependencies are not local anymore, and approximate inference is required.

We use the algorithm described in section 2 (figure 2) to infer the state. In the context of the semantics-text correspondence model, as we discussed above, semantics $\boldsymbol{m}$ defines the subset of admissible world states. In order to use the algorithm, we need to understand how the conditional probabilities of the form $P(\boldsymbol{m}'|\boldsymbol{m})$ are computed, as they play the key role in the inference procedure (see equation (2)). If there is a contradiction ($\boldsymbol{m}' \perp \boldsymbol{m}$) then $P(\boldsymbol{m}'|\boldsymbol{m}) = 0$, conversely, if $\boldsymbol{m}'$ is subsumed by $\boldsymbol{m}$ ($\boldsymbol{m} \rightarrow \boldsymbol{m}'$) then this probability is 1. Otherwise, $P(\boldsymbol{m}'|\boldsymbol{m})$ equals the probability of new assignments $\wedge_{q=1}^{|\boldsymbol{m}'\setminus\boldsymbol{m}|}(s_{n'_q,f'_q}^{(t'_q)} = v'_q)$ (defined by $\boldsymbol{m}'\setminus\boldsymbol{m}$) conditioned on the previously fixed values of $\boldsymbol{s}$ (given by $\boldsymbol{m}$). Summarizing, when predicting the most likely semantics $\hat{\boldsymbol{m}}_j$ (line 4), for each span the decoder weighs alternatives of either (1) aligning this span to the previously induced meaning $\boldsymbol{m}^\star$, or (2) aligning it to a new field and paying the cost of generation of its value.

The exact computation of the most probable semantics (line 4 of the algorithm) is intractable, and we have to resort to an approximation. Instead of predicting the most probable semantics $\hat{\boldsymbol{m}}_j$ we search for the most probable pair $(\hat{\boldsymbol{a}}_j, \hat{\boldsymbol{m}}_j)$, thus assuming that the probability mass is mostly concentrated on a single alignment. The alignment $\boldsymbol{a}_j$

[5]For simplicity, we assume here that all the examples are unlabeled.

is then discarded and not used in any other computations. Though the most likely alignment $\hat{\boldsymbol{a}}_j$ for a fixed semantic representation $\hat{\boldsymbol{m}}_j$ can be found efficiently using a Viterbi algorithm, computing the most probable pair $(\hat{\boldsymbol{a}}_j, \hat{\boldsymbol{m}}_j)$ is still intractable. We use a modification of the beam search algorithm, where we keep a set of candidate meanings (partial semantic representations) and compute an alignment for each of them using a form of the Viterbi algorithm.

As soon as the meaning representations $\boldsymbol{m}^\star$ are inferred, we find ourselves in the set-up studied in (Liang et al., 2009): the state $\boldsymbol{s}$ is no longer latent and we can run efficient inference on the E-step. Though some fields of the state $\boldsymbol{s}$ may still not be specified by $\boldsymbol{m}^\star$, we prohibit utterances from aligning to these non-specified fields.

On the M-step of EM the parameters are estimated as proportional to the expected marginal counts computed on the E-step. We smooth the distributions of values for numerical fields with convolution smoothing equivalent to the assumption that the fields are affected by distortion in the form of a two-sided geometric distribution with the success rate parameter equal to 0.67. We use add-0.1 smoothing for all the remaining multinomial distributions.

4 Empirical Evaluation

In this section, we consider the semi-supervised set-up, and present evaluation of our approach on on the problem of aligning weather forecast reports to the formal representation of weather.

4.1 Experiments

To perform the experiments we used a subset of the weather dataset introduced in (Liang et al., 2009). The original dataset contains 22,146 texts of 28.7 words on average, there are 12 types of records (predicates) and 36.0 records per forecast on average. We randomly chose 100 texts along with their world states to be used as the labeled data.[6] To produce groups of non-contradictory texts we have randomly selected a subset of weather states, represented them in a visual form (icons accompanied by numerical and symbolic parameters) and then manually annotated these illustrations. These newly-produced forecasts, when combined with the original texts, resulted in 259 groups of non-contradictory texts (650 texts, 2.5 texts per group). An example of such a group is given in figure 1.

The dataset is relatively noisy: there are inconsistencies due to annotation mistakes (e.g., number distortions), or due to different perception of the weather by the annotators (e.g., expressions such as 'warm' or 'cold' are subjective). The overlap between the verbalized fields in each group was estimated to be below 35%. Around 60% of fields are mentioned only in a single forecast from a group, consequently, the texts cannot be regarded as paraphrases of each other.

The test set consists of 150 texts, each corresponding to a different weather state. Note that during testing we no longer assume that documents share the state, we treat each document in isolation. We aimed to preserve approximately the same proportion of new and original examples as we had in the training set, therefore, we combined 50 texts originally present in the weather dataset with additional 100 newly-produced texts. We annotated these 100 texts by aligning each line to one or more records,[7] whereas for the original texts the alignments were already present. Following Liang et al. (2009) we evaluate the models on how well they predict these alignments.

When estimating the model parameters, we followed the training regime prescribed in (Liang et al., 2009). Namely, 5 iterations of EM with a basic model (with no segmentation or coherence modeling), followed by 5 iterations of EM with the model which generates fields independently and, at last, 5 iterations with the full model. Only then, in the semi-supervised learning scenarios, we added unlabeled data and ran 5 additional iterations of EM.

Instead of prohibiting records from crossing punctuation, as suggested by Liang et al. (2009), in our implementation we disregard the words not attached to specific fields (attached to the null-field, see section 3.1) when computing spans of records. To speed-up training, only a single record of each type is allowed to be generated when running inference for unlabeled examples on the E-

[6] In order to distinguish from completely unlabeled examples, we refer to examples labeled with world states as *labeled* examples. Note though that the alignments are not observable even for these labeled examples. Similarly, we call the models trained from this data *supervised* though full supervision was not available.

[7] The text was automatically tokenized and segmented into lines, with line breaks at punctuation characters. Information about the line breaks is not used during learning and inference.

	P	R	F_1
Supervised BL	63.3	52.9	57.6
Semi-superv BL	68.8	69.4	69.1
Semi-superv, non-contr	**78.8**	**69.5**	**73.9**
Supervised UB	69.4	88.6	77.9

Table 1: Results (precision, recall and F_1) on the weather forecast dataset.

value	top words
0-25	clear, small, cloudy, gaps, sun
25-50	clouds, increasing, heavy, produce, could
50-75	cloudy, mostly, high, cloudiness, breezy
75-100	amounts, rainfall, inch, new, possibly

Table 2: Top 5 words in the word distribution for field *mode* of record *sky cover*, function words and punctuation are omitted.

step of the EM algorithm, as it significantly reduces the search space. Similarly, though we preserved all records which refer to the first time period, for other time periods we removed all the records which declare that the corresponding event (e.g., rain or snowfall) is not expected to happen. This preprocessing results in the oracle recall of 93%.

We compare our approach (*Semi-superv, non-contr*) with two baselines: the basic supervised training on 100 labeled forecasts (*Supervised BL*) and with the semi-supervised training which disregards the non-contradiction relations (*Semi-superv BL*). The learning regime, the inference procedure and the texts for the semi-supervised baseline were identical to the ones used for our approach, the only difference is that all the documents were modeled as independent. Additionally, we report the results of the model trained with all the 750 texts labeled (*Supervised UB*), its scores can be regarded as an upper bound on the results of the semi-supervised models. The results are reported in table 1.

4.2 Discussion

Our training strategy results in a substantially more accurate model, outperforming both the supervised and semi-supervised baselines. Surprisingly, its precision is higher than that of the model trained on 750 labeled examples, though admittedly it is achieved at a very different recall level. The estimation of the model with our approach takes around one hour on a standard desktop PC, which is comparable to 40 minutes required to train the semi-supervised baseline.

In these experiments, we consider the problem of predicting alignment between text and the corresponding observable world state. The direct evaluation of the meaning recognition (i.e. semantic parsing) accuracy is not possible on this dataset, as the data does not contain information which fields are discussed. Even if it would provide this information, the documents do not verbalize the state at the necessary granularity level to predict the field values. For example, it is not possible to decide to which bucket of the field *sky cover* the expression 'cloudy' refers to, as it has a relatively uniform distribution across 3 (out of 4) buckets. The problem of predicting text-meaning alignments is interesting in itself, as the extracted alignments can be used in training of a statistical generation system or information extractors, but we also believe that evaluation on this problem is an appropriate test for the relative comparison of the semantic analyzers' performance. Additionally, note that the success of our weakly-supervised scenario indirectly suggests that the model is sufficiently accurate in predicting semantics of an unlabeled text, as otherwise there would be no useful information passed in between semantically overlapping documents during learning and, consequently, no improvement from sharing the state.[8]

To confirm that the model trained by our approach indeed assigns new words to correct fields and records, we visualize top words for the field characterizing sky cover (table 2). Note that the words "sun", "cloudiness" or "gaps" were not appearing in the labeled part of the data, but seem to be assigned to correct categories. However, correlation between rain and overcast, as also noted in (Liang et al., 2009), results in the wrong assignment of the rain-related words to the field value corresponding to very cloudy weather.

5 Related Work

Probably the most relevant prior work is an approach to bootstrapping lexical choice of a generation system using a corpus of alternative pas-

[8] We conducted preliminary experiments on synthetic data generated from a random semantic-correspondence model. Our approach outperformed the baselines both in predicting 'text'-state correspondence and in the F_1 score on the predicted set of field assignments ('text meanings').

sages (Barzilay and Lee, 2002), however, in their work all the passages were annotated with unaligned semantic expressions. Also, they assumed that the passages are paraphrases of each other, which is stronger than our non-contradiction assumption. Sentence and text alignment has also been considered in the related context of paraphrase extraction (see, e.g., (Dolan et al., 2004; Barzilay and Lee, 2003)) but this prior work did not focus on inducing or learning semantic representations. Similarly, in information extraction, there have been approaches for pattern discovery using comparable monolingual corpora (Shinyama and Sekine, 2003) but they generally focused only on discovery of a single pattern from a pair of sentences or texts.

Radev (2000) considered types of potential relations between documents, including contradiction, and studied how this information can be exploited in NLP. However, this work considered primarily multi-document summarization and question answering problems.

Another related line of research in machine learning is clustering or classification with constraints (Basu et al., 2004), where supervision is given in the form of constraints. Constraints declare which pairs of instances are required to be assigned to the same class (or required to be assigned to different classes). However, we are not aware of any previous work that generalized these methods to structured prediction problems, as trivial equality/inequality constraints are probably too restrictive, and a notion of consistency is required instead.

6 Summary and Future Work

In this work we studied the use of weak supervision in the form of non-contradictory relations between documents in learning semantic representations. We argued that this type of supervision encodes information which is hard to discover in an unsupervised way. However, exact inference for groups of documents with overlapping semantic representation is generally prohibitively expensive, as the shared latent semantics introduces nonlocal dependences between semantic representations of individual documents. To combat it, we proposed a simple iterative inference algorithm. We showed how it can be instantiated for the semantics-text correspondence model (Liang et al., 2009) and evaluated it on a dataset of weather forecasts. Our approach resulted in an improvement over the scores of both the supervised baseline and of the traditional semi-supervised learning.

There are many directions we plan on investigating in the future for the problem of learning semantics with non-contradictory relations. A promising and challenging possibility is to consider models which induce full semantic representations of meaning. Another direction would be to investigate purely unsupervised set-up, though it would make evaluation of the resulting method much more complex. One potential alternative would be to replace the initial supervision with a set of posterior constraints (Graca et al., 2008) or generalized expectation criteria (McCallum et al., 2007).

Acknowledgements

The authors acknowledge the support of the Excellence Cluster on Multimodal Computing and Interaction (MMCI). Thanks to Alexandre Klementiev, Alexander Koller, Manfred Pinkal, Dan Roth, Caroline Sporleder and the anonymous reviewers for their suggestions, and to Percy Liang for answering questions about his model.

References

Regina Barzilay and Lillian Lee. 2002. Bootstrapping lexical choice via multiple-sequence alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 164–171.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the Conference on Human Language Technology and North American chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Sugatu Basu, Arindam Banjeree, and Raymond Mooney. 2004. Active semi-supervision for pairwise constrained clustering. In *Proc. of the SIAM International Conference on Data Mining (SDM)*, pages 333–344.

A. Blum and T. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, pages 209–214.

Xavier Carreras and Lluis Marquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of CoNLL-2005*, Ann Arbor, MI USA.

David L. Chen and Raymond L. Mooney. 2008. Learning to sportcast: A test of grounded language acquisition. In *Proc. of International Conference on Machine Learning*, pages 128–135.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithms. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

P. Diaconis and B. Efron. 1983. Computer-intensive methods in statistics. *Scientific American*, pages 116–130.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the Conference on Computational Linguistics (COLING)*, pages 350–356.

Ruifang Ge and Raymond J. Mooney. 2005. A statistical semantic parser that integrates syntax and semantics. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CONLL-05)*, Ann Arbor, Michigan.

Joao Graca, Kuzman Ganchev, and Ben Taskar. 2008. Expectation maximization and posterior constraints. *Advances in Neural Information Processing Systems 20 (NIPS)*.

Zellig Harris. 1968. *Mathematical structures of language*. Wiley.

Rohit J. Kate and Raymond J. Mooney. 2007. Learning language semantics from ambigous supervision. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 895–900.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proc. of the Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Andrew McCallum, Gideon Mann, and Gregory Druck. 2007. Generalized expectation criteria. Technical Report TR 2007-60, University of Massachusetts, Amherst, MA.

Raymond J. Mooney. 2007. Learning for semantic parsing. In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 982–991.

Kevin P. Murphy, Yair Weiss, and Michael I. Jordan. 1999. Loopy belief propagation for approximate inference: An empirical study. In *Proc. of Uncertainty in Artificial Intelligence (UAI)*, pages 467–475.

Judea Pearl. 1982. Reverend bayes on inference engines: A distributed hierarchical approach. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*, pages 133–136.

Hoifung Poon and Pedro Domingos. 2009. Unsupervised semantic parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, (EMNLP-09)*.

Dragomir Radev. 2000. A common theory of information fusion from multiple text sources step one: Cross-document structure. In *1st SIGdial Workshop on Discourse and Dialogue*, pages 74–83.

Yusuke Shinyama and Satoshi Sekine. 2003. Paraphrase acquisition for information extraction. In *Proceedings of Second International Workshop on Paraphrasing (IWP2003)*, pages 65–71.

Benjamin Snyder and Regina Barzilay. 2007. Database-text alignment via structured multilabel classification. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-05)*, pages 1713–1718.

J. Weeds and W. Weir. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475.

Luke Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammar. In *Proceedings of the Twenty-first Conference on Uncertainty in Artificial Intelligence*, Edinburgh, UK, August.