Two Discourse Driven Language Models for Semantics

Haoruo Peng and **Dan Roth**
University of Illinois, Urbana-Champaign
Urbana, IL, 61801
{hpeng7,danr}@illinois.edu

Abstract

Natural language understanding often requires deep semantic knowledge. Expanding on previous proposals, we suggest that some important aspects of semantic knowledge can be modeled as a language model if done at an appropriate level of abstraction. We develop two distinct models that capture semantic frame chains and discourse information while abstracting over the specific mentions of predicates and entities. For each model, we investigate four implementations: a "standard" N-gram language model and three discriminatively trained "neural" language models that generate embeddings for semantic frames. The quality of the semantic language models (SemLM) is evaluated both intrinsically, using perplexity and a narrative cloze test and extrinsically – we show that our SemLM helps improve performance on semantic natural language processing tasks such as co-reference resolution and discourse parsing.

1 Introduction

Natural language understanding often necessitates deep semantic knowledge. This knowledge needs to be captured at multiple levels, from words to phrases, to sentences, to larger units of discourse. At each level, capturing meaning frequently requires context sensitive abstraction and disambiguation, as shown in the following example (Winograd, 1972):

Ex.1 *[Kevin] was **robbed** by [Robert]. [He] was **arrested** by the police.*

Ex.2 *[Kevin] was **robbed** by [Robert]. [He] was **rescued** by the police.*

In both cases, one needs to resolve the pronoun "he" to either "Robert" or "Kevin". To make the correct decisions, one needs to know that the subject of "rob" is more likely than the object of "rob" to be the object of "arrest" while the object of "rob" is more likely to be the object of "rescue". Thus, beyond understanding individual predicates (e.g., at the semantic role labeling level), there is a need to place them and their arguments in a global context.

However, just modeling semantic frames is not sufficient; consider a variation of Ex.1:

Ex.3 *Kevin was **robbed** by Robert, but the police mistakenly **arrested** him.*

In this case, "him" should refer to "Kevin" as the discourse marker "but" reverses the meaning, illustrating that it is necessary to take discourse markers into account when modeling semantics.

In this paper we propose that these aspects of semantic knowledge can be modeled as a *Semantic Language Model* (SemLM). Just like the "standard" syntactic language models (LM), we define a basic vocabulary, a finite representation language, and a prediction task, which allows us to model the distribution over the occurrence of elements in the vocabulary as a function of their (well-defined) context. In difference from syntactic LMs, we represent natural language at a higher level of semantic abstraction, thus facilitating modeling deep semantic knowledge.

We propose two distinct discourse driven language models to capture semantics. In our first semantic language model, the *Frame-Chain SemLM*, we model all semantic frames and discourse markers in the text. Each document is viewed as a single chain of semantic frames and discourse markers. Moreover, while the vocabulary of discourse markers is rather small, the number of different surface form semantic frames that could appear in the text is very large. To achieve a better level of abstraction, we disambiguate semantic frames and map them to their PropBank/FrameNet represen-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 290–300,
Berlin, Germany, August 7-12, 2016.

tation. Thus, in Ex.3, the resulting frame chain is "rob.01 — but — arrest.01" ("01" indicates the predicate sense).

Our second semantic language model is called *Entity-Centered SemLM*. Here, we model a sequence of semantic frames and discourse markers involved in a specific co-reference chain. For each co-reference chain in a document, we first extract semantic frames corresponding to each co-referent mention, disambiguate them as before, and then determine the discourse markers between these frames. Thus, each unique frame contains both the disambiguated predicate and the argument label of the mention. In Ex.3, the resulting sequence is "rob.01#obj — but — arrest.01#obj" (here "obj" indicates the argument label for "Kevin" and "him" respectively). While these two models capture somewhat different semantic knowledge, we argue later in the paper that both models can be induced at high quality, and that they are suitable for different NLP tasks.

For both models of SemLM, we study four language model implementations: N-gram, skip-gram (Mikolov et al., 2013b), continuous bag-of-words (Mikolov et al., 2013a) and log-bilinear language model (Mnih and Hinton, 2007). Each model defines its own prediction task. In total, we produce eight different SemLMs. Except for N-gram model, others yield embeddings for semantic frames as they are neural language models.

In our empirical study, we evaluate both the quality of all SemLMs and their application to co-reference resolution and shallow discourse parsing tasks. Following the traditional evaluation standard of language models, we first use perplexity as our metric. We also follow the script learning literature (Chambers and Jurafsky, 2008b; Chambers and Jurafsky, 2009; Rudinger et al., 2015) and evaluate on the narrative cloze test, i.e. randomly removing a token from a sequence and test the system's ability to recover it. We conduct both evaluations on two test sets: a hold-out dataset from the New York Times Corpus and gold sequence data (for frame-chain SemLMs, we use PropBank (Kingsbury and Palmer, 2002); for entity-centered SemLMs, we use Ontonotes (Hovy et al., 2006)). By comparing the results on these test sets, we show that we do not incur noticeable degradation when building SemLMs using preprocessing tools. Moreover, we show that SemLMs improves the performance of co-reference resolution, as well as that of predicting the sense of discourse connectives for both explicit and implicit ones.

The main contributions of our work can be summarized as follows: 1) The design of two novel discourse driven Semantic Language models, building on text abstraction and neural embeddings; 2) The implementation of high quality SemLMs that are shown to improve state-of-the-art NLP systems.

2 Related Work

Our work is related to script learning. Early works (Schank and Abelson, 1977; Mooney and DeJong, 1985) tried to construct knowledge bases from documents to learn scripts. Recent work focused on utilizing statistical models to extract high-quality scripts from large amounts of data (Chambers and Jurafsky, 2008a; Bejan, 2008; Jans et al., 2012; Pichotta and Mooney, 2014; Granroth-Wilding et al., 2015; Pichotta and Mooney, 2016). Other works aimed at learning a collection of structured events (Chambers, 2013; Cheung et al., 2013; Cheung et al., 2013; Balasubramanian et al., 2013; Bamman and Smith, 2014; Nguyen et al., 2015), and several works have employed neural embeddings (Modi and Titov, 2014b; Modi and Titov, 2014a; Frermann et al., 2014; Titov and Khoddam, 2015).

In our work, the semantic sequences in the entity-centered SemLMs are similar to narrative schemas (Chambers and Jurafsky, 2009). However, we differ from them in the following aspects: 1) script learning does not generate a probabilistic model on semantic frames[1]; 2) script learning models semantic frame sequences incompletely as they do not consider discourse information; 3) works in script learning rarely show applications to real NLP tasks.

Some prior works have used scripts-related ideas to help improve NLP tasks (Irwin et al., 2011; Rahman and Ng, 2011; Peng et al., 2015b). However, since they use explicit script schemas either as features or constraints, these works suffer from data sparsity problems. In our work, the SemLM abstract vocabulary ensures a good coverage of frame semantics.

[1]Some works may utilize a certain probabilistic framework, but they mainly focus on generating high-quality frames by filtering.

Table 1: **Comparison of vocabularies between frame-chain (FC) and entity-centered (EC) SemLMs.** "F-Sen" stands for frames with predicate sense information while "F-Arg" stands for frames with argument role label information; "Conn" means discourse marker and "Per" means period. "Seq/Doc" represents the number of sequence per document.

	F-Sen	F-Arg	Conn	Per	Seq/Doc
FC	YES	NO	YES	YES	Single
EC	YES	YES	YES	NO	Multiple

3 Two Models for SemLM

In this section, we describe how we capture sequential semantic information consisted of semantic frames and discourse markers as semantic units (i.e. the vocabulary).

3.1 Semantic Frames and Discourse Markers

Semantic Frames A semantic frame is composed of a predicate and its corresponding argument participants. Here we require the predicate to be disambiguated to a specific sense, and we need a certain level of abstraction of arguments so that we can assign abstract labels. The design of PropBank frames (Kingsbury and Palmer, 2002) and FrameNet frames (Baker et al., 1998) perfectly fits our needs. They both have a limited set of frames (in the scale of thousands) and each frame can be uniquely represented by its predicate sense. These frames provide a good level of generalization as each frame can be instantiated into various surface forms in natural texts. We use these frames as part of our vocabulary for SemLMs. Formally, we use the notation f to represent a frame. Also, we denote fa $\triangleq$ f#Arg when referring to an argument role label (Arg) inside a frame (f).

Discourse Markers We use discourse markers (connectives) to model discourse relationships between frames. There is only a limited number of unique discourse markers, such as *and, but, however*, etc. We get the full list from the Penn Discourse Treebank (Prasad et al., 2008) and include them as part of our vocabulary for SemLMs. Formally, we use *dis* to denote the discourse marker. Note that discourse relationships can exist without an explicit discourse marker, which is also a challenge for discourse parsing. Since we cannot reliably identify implicit discourse relationships, we only consider explicit ones here. More importantly, discourse markers are associated with arguments (Wellner and Pustejovsky, 2007) in text (usually two sentences/clauses, sometimes one). We only add a discourse marker in the semantic sequence when its corresponding arguments contain semantic frames which belong to the same semantic sequence. We call them *frame-related discourse markers*. Details on generating semantic frames and discourse markers to form semantic sequences are discussed in Sec. 5.

3.2 Frame-Chain SemLM

For frame-chain SemLM, we model all semantic frames and discourse markers in a document. We form the semantic sequence by first including all semantic frames in the order they appear in the text: $[\mathrm{f}_1, \mathrm{f}_2, \mathrm{f}_3, \ldots]$. Then we add *frame-related discourse markers* into the sequence by placing them in their order of appearance. Thus we get a sequence like $[\mathrm{f}_1, \mathrm{dis}_1, \mathrm{f}_2, \mathrm{f}_3, \mathrm{dis}_2, \ldots]$. Note that discourse markers do not necessarily exist between all semantic frames. Additionally, we treat the *period* symbol as a special discourse marker, denoted by "o". As some sentences contain more than one semantic frame (situations like clauses), we get the final semantic sequence like this:

$$[\mathrm{f}_1, \mathrm{dis}_1, \mathrm{f}_2, \mathrm{o}, \mathrm{f}_3, \mathrm{o}, \mathrm{dis}_2, \ldots, \mathrm{o}]$$

3.3 Entity-Centered SemLM

We generate semantic sequences according to co-reference chains for entity-centered SemLM. From co-reference resolution, we can get a sequence like $[\mathrm{m}_1, \mathrm{m}_2, \mathrm{m}_3, \ldots]$, where mentions appear in the order they occur in the text. Each mention can be matched to an argument inside a semantic frame. Thus, we replace each mention with its argument label inside a semantic frame, and get $[\mathrm{fa}_1, \mathrm{fa}_2, \mathrm{fa}_3, \ldots]$. We then add discourse markers exactly in they way we do for frame-chain SemLM, and get the following sequence:

$$[\mathrm{fa}_1, \mathrm{dis}_1, \mathrm{fa}_2, \mathrm{fa}_3, \mathrm{dis}_2, \ldots]$$

The comparison of vocabularies between frame-chain and entity-centered SemLMs is summarized in Table 1.

4 Implementations of SemLM

In this work, we experiment with four language model implementations: N-gram (NG), Skip-Gram (SG), Continuous Bag-of-Words (CBOW) and Log-bilinear (LB) language model. For ease

of explanation, we assume that a semantic unit sequence is $s = [w_1, w_2, w_3, \ldots, w_k]$.

4.1 N-gram Model

For an n-gram model, we predict each token based on its $n-1$ previous tokens, i.e. we directly model the following conditional probability (in practice, we choose $n = 3$, Tri-gram (TRI)):

$$p(w_{t+2}|w_t, w_{t+1}).$$

Then, the probability of the sequence is

$$p(s) = p(w_1)p(w_2|w_1)\prod_{t=1}^{k-2} p(w_{t+2}|w_t, w_{t+1}).$$

To compute $p(w_2|w_1)$ and $p(w_1)$, we need to back off from Tri-gram to Bi-gram and Uni-gram.

4.2 Skip-Gram Model

The SG model was proposed in Mikolov et al. (2013b). It uses a token to predict its context, i.e. we model the following conditional probability:

$$p(c \in c(w_t)|w_t, \theta).$$

Here, $c(w_t)$ is the context for w_t and θ denotes the learned parameters which include neural network states and embeddings. Then the probability of the sequence is computed as

$$\prod_{t=1}^{k} \prod_{c \in c(w_t)} p(c|w_t, \theta).$$

4.3 Continuous Bag-of-Words Model

In contrast to skip-gram, CBOW (Mikolov et al., 2013a) uses context to predict each token, i.e. we model the following conditional probability:

$$p(w_t|c(w_t), \theta).$$

In this case, the probability of the sequence is

$$\prod_{t=1}^{k} p(w_t|c(w_t), \theta).$$

4.4 Log-bilinear Model

LB was introduced in Mnih and Hinton (2007). Similar to CBOW, it also uses context to predict each token. However, LB associates a token with three components instead of just one vector: a target vector v(w), a context vector v'(w) and a bias b(w). So, the conditional probability becomes:

$$p(w_t|c(w_t)) = \frac{\exp(v(w_t)^\intercal u(c(w_t)) + b(w_t))}{\sum_{w \in \mathcal{V}} \exp(v(w)^\intercal u(c(w_t)) + b(w))}.$$

Here, $\mathcal{V}$ denotes the vocabulary and we define $u(c(w_t)) = \sum_{c_i \in c(w_t)} q_i \odot v'(c_i)$. Note that $\odot$ represents element-wise multiplication and q_i is a vector that depends only on the position of a token in the context, which is a also a model parameter.

So, the overall sequence probability is

$$\prod_{t=1}^{k} p(w_t|c(w_t)).$$

5 Building SemLMs from Scratch

In this section, we explain how we build SemLMs from un-annotated plain text.

5.1 Dataset and Preprocessing

Dataset We use the New York Times Corpus[2] (from year 1987 to 2007) for training. It contains a bit more than 1.8M documents in total.

Preprocessing We pre-process all documents with semantic role labeling (Punyakanok et al., 2004) and part-of-speech tagger (Roth and Zelenko, 1998). We also implement the explicit discourse connective identification module in shallow discourse parsing (Song et al., 2015). Additionally, we utilize within document entity coreference (Peng et al., 2015a) to produce coreference chains. To obtain all annotations, we employ the Illinois NLP tools[3].

5.2 Semantic Unit Generation

FrameNet Mapping We first directly derive semantic frames from semantic role labeling annotations. As the Illinois SRL package is built upon PropBank frames, we do a mapping to FrameNet frames via VerbNet senses (Schuler, 2005), thus achieving a higher level of abstraction. The mapping file[4] defines deterministic mappings. However, the mapping is not complete and there are remaining PropBank frames. Thus, the generated vocabulary for SemLMs contains both PropBank and FrameNet frames. For example, "place" and

[2]https://catalog.ldc.upenn.edu/LDC2008T19

[3]http://cogcomp.cs.illinois.edu/page/software/

[4]http://verbs.colorado.edu/verb-index/fn/vn-fn.xml

"put" with the VerbNet sense id "9.1-2" are converted to the same FrameNet frame "Placing".

Augmenting to Verb Phrases We apply three heuristic modifications to augment semantic frames defined in Sec. 3.1: 1) if a preposition immediately follows a predicate, we append the preposition to the predicate e.g. "take over"; 2) if we encounter the semantic role label AM-PRD which indicates a secondary predicate, we also append this secondary predicate to the main predicate e.g. "be happy"; 3) if we see the semantic role label AM-NEG which indicates negation, we append "not" to the predicate e.g. "not like". These three augmentations can co-exist and they allow us to model more fine-grained semantic frames.

Verb Compounds We have observed that if two predicates appear very close to each other, e.g. "eat and drink", "decide to buy", they actually represent a unified semantic meaning. Thus, we construct compound verbs to connect them together. We apply the rule that if the gap between two predicates is less than two tokens, we treat them as a unified semantic frame defined by the conjunction of the two (augmented) semantic frames, e.g. "eat.01-drink.01" and "decide.01-buy.01".

Argument Labels for Co-referent Mentions To get the argument role label information for co-referent mentions, we need to match each mention to its corresponding semantic role labeling argument. If a mention head is inside an argument, we regard it as a match. We do not consider singleton mentions.

Vocabulary Construction After generating all semantic units for (augmented and compounded) semantic frames and discourse markers, we merge them together as a tentative vocabulary. In order to generate a sensible SemLM, we filter out rare tokens which appear less than 20 times in the data. We add the Unknown token (UNK) and End-of-Sequence token (EOS) to the eventual vocabulary.

Statistics on the eventual SemLM vocabularies and semantic sequences are shown in Table 2. We also compare frame-chain and entity-centered SemLMs to the usual syntactic language model setting. The statistics in Table 2 shows that they are comparable both in vocabulary size and in the total number of tokens for training. Moreover, entity-centered SemLMs have shorter sequences then frame-chain SemLMs. We also provide several examples of high-frequency augmented compound semantic frames in our generated SemLM vocabularies. All are very intuitive:

want.01-know.01, agree.01-pay.01, try.01-get.01, decline.02-comment.01, wait.01-see.01, make.02-feel.01, want.01(not)-give.08(up)

Table 2: **Statistics on SemLM vocabularies and sequences.** "F-s" stands for single frame while "F-c" stands for compound frame; "Conn" means discourse marker. "#seq" is the number of sequences, and "#token" is the total number of tokens (semantic units). We also compute the average token in a sequence i.e. "#t/s". We compare frame-chain (FC) and entity-centered (EC) SemLMs to the usual syntactic language model setting i.e. "LM".

	Vocabulary Size			Sequence Size		
	F-s	F-c	Conn	#seq	#token	#t/s
FC	14857	7269	44	1.2M	25.4M	21
EC	8758	2896	44	3.4M	18.6M	5
LM	∼20k			∼3M	∼38M	10-15

5.3 Language Model Training

NG We implement the N-gram model using the SRILM toolkit (Stolcke, 2002). We also employ the well-known KneserNey Smoothing (Kneser and Ney, 1995) technique.

SG & CBOW We utilize the word2vec package to implement both SG and CBOW. In practice, we set the context window size to be 10 for SG while set the number as 5 for CBOW (both are usual settings for syntactic language models). We generate 300-dimension embeddings for both models.

LB We use the OxLM toolkit (Paul et al., 2014) with Noise-Constrastive Estimation (Gutmann and Hyvarinen, 2010) for the LB model. We set the context window size to 5 and produce 150-dimension embeddings.

6 Evaluation

In this section, we first evaluate the quality of SemLMs through perplexity and a narrative cloze test. More importantly, we show that the proposed SemLMs can help improve the performance of co-reference resolution and shallow discourse parsing. This further proves that we successfully capture semantic sequence information which can potentially benefit a wide range of semantic related NLP tasks.

We have designed two models for SemLM: *frame-chain* (**FC**) and *entity-centered* (**EC**). By training on both types of sequences respectively, we implement four different language models:

TRI, SG, CBOW, LB. We focus the evaluation efforts on these eight SemLMs.

6.1 Quality Evaluation of SemLMs

Datasets We use three datasets. We first randomly sample 10% of the New York Times Corpus documents (roughly two years of data), denoted the *NYT Hold-out Data*. All our SemLMs are trained on the remaining NYT data and tested on this hold-out data. We generate semantic sequences for the training and test data using the methodology described in Sec. 5.

We use PropBank data with gold frame annotations as another test set. In this case, we only generate frame-chain SemLM sequences by applying semantic unit generation techniques on gold frames, as described in Sec 5.2. When we test on *Gold PropBank Data with Frame Chains*, we use frame-chain SemLMs trained from all NYT data.

Similarly, we use Ontonotes data (Hovy et al., 2006) with gold frame and co-reference annotations as the third test set, *Gold Ontonotes Data with Coref Chains*. We only generate entity-centered SemLMs by applying semantic unit generation techniques on gold frames and gold co-reference chains, as described in Sec 5.2.

Baselines We use Uni-gram (**UNI**) and Bi-gram (**BG**) as two language model baselines. In addition, we use the point-wise mutual information (PMI) for token prediction. Essentially, PMI scores each pair of tokens according to their co-occurrences. It predicts a token in the sequence by choosing the one with the highest total PMI with all other tokens in the sequence. We use the ordered PMI (**OP**) as our baseline, which is a variation of PMI by considering asymmetric counting (Jans et al., 2012).

6.1.1 Perplexity

As SemLMs are language models, it is natural to evaluate the perplexity, which is a measurement of how well a language model can predict sequences.

Results for SemLM perplexities are presented in Table 3. They are computed without considering end token (EOS). We apply tri-gram Kneser-Ney Smoothing to CBOW, SG and LB. LB consistently shows the lowest perplexities for both frame-chain and entity-centered SemLMs across all test sets. Similar to syntactic language models, perplexities are fast decreasing from UNI, BI to TRI. Also, CBOW and SG have very close perplexity results which indicate that their language modeling abilities are at the same level.

Table 3: **Perplexities for SemLMs.** UNI, BG, TRI, CBOW, SG, LB are different language model implementations while "FC" and "EC" stand for the two SemLM models studied, respectively. "FC-FM" and "EC-FM" indicate that we removed the "FrameNet Mapping" step (Sec. 5.2). LB consistently produces the lowest perplexities for both frame-chain and entity-centered SemLMs.

	Baselines		SemLMs			
	UNI	BG	TRI	CBOW	SG	LB
NYT Hold-out Data						
FC	952.1	178.3	119.2	115.4	114.1	**108.5**
EC	914.7	154.4	114.9	111.8	113.8	**109.7**
Gold PropBank Data with Frame Chains						
FC-FM	992.9	213.7	139.1	135.6	128.4	121.8
FC	970.0	191.2	132.7	126.4	123.5	**115.4**
Gold Ontonotes Data with Coref Chains						
EC-FM	956.4	187.7	121.1	115.6	117.2	113.7
EC	923.8	163.2	120.5	113.7	115.0	**109.3**

We can compare the results of our frame-chain SemLM on *NYT Hold-out Data* and *Gold PropBank Data with Frame Chains*, and our entity-centered SemLM on *NYT Hold-out Data* and *Gold Ontonotes Data with Coref Chains*. While we see differences in the results, the gap is narrow and the relative ranking of different SemLMs does not change. This indicates that the automatic SRL and Co-reference annotations added some noise but, more importantly, that the resulting SemLMs are robust to this noise as we still retain the language modeling ability for all methods.

Additionally, our ablation study removes the "FrameNet Mapping" step in Sec. 5.2 ("FC-FM" and "EC-FM" rows), resulting in only using PropBank frames in the vocabulary. The increase in perplexities shows that "FrameNet Mapping" does produce a higher level of abstraction, which is useful for language modeling.

6.1.2 Narrative Cloze Test

We follow the Narrative Cloze Test idea used in script learning (Chambers and Jurafsky, 2008b; Chambers and Jurafsky, 2009). As Rudinger et al. (2015) points out, the narrative cloze test can be regarded as a language modeling evaluation. In the narrative cloze test, we randomly choose and remove one token from each semantic sequence in the test set. We then use language models to predict the missing token and evaluate the correctness. For all SemLMs, we use the conditional probabilities defined in Sec. 4 to get token predictions. We also use ordered PMI as an additional baseline. The narrative cloze test is conducted on

Table 4: **Narrative cloze test results for SemLMs.** UNI, BG, TRI, CBOW, SG, LB are different language model implementations while "FC" and "EC" stand for our two SemLM models, respectively. "FC-FM" and "EC-FM" mean that we remove the FrameNet mappings. "w/o DIS" indicates the removal of discourse makers in SemLMs. "Rel-Impr" indicates the relative improvement of the best performing SemLM over the strongest baseline. We evaluate on two metrics: mean reciprocal rank (MRR)/recall at 30 (Recall@30). LB outperforms other methods for both frame-chain and entity-centered SemLMs.

	Baselines			SemLMs				Rel-Impr
	OP	UNI	BG	TRI	CBOW	SG	LB	
MRR								
NYT Hold-out Data								
FC	0.121	0.236	0.225	0.249	0.242	0.247	**0.276**	8.5%
EC	0.126	0.235	0.210	0.242	0.249	0.249	**0.261**	5.9%
EC w/o DIS	0.092	0.191	0.188	0.212	0.215	0.216	**0.227**	18.8%
Rudinger et al. (2015)*	0.083	0.186	0.181	—	—	—	**0.223**	19.9%
Gold PropBank Data with Frame Chains								
FC	0.106	0.215	0.212	0.232	0.228	0.229	**0.254**	18.1%
FC-FM	0.098	0.201	0.204	0.223	0.218	0.220	0.243	———
Gold Ontonotes Data with Coref Chains								
EC	0.122	0.228	0.213	0.239	0.247	0.246	**0.257**	12.7%
EC-FM	0.109	0.215	0.208	0.230	0.237	0.239	0.254	———
Recall@30								
NYT Hold-out Data								
FC	33.2	46.8	45.3	47.3	46.6	47.5	**55.4**	18.4%
EC	29.4	43.7	41.6	44.8	46.5	46.6	**52.0**	19.0%
Gold PropBank Data with Frame Chains								
FC	26.3	39.5	38.1	45.5	43.6	43.8	**53.9**	36.5%
FC-FM	24.4	37.3	37.3	42.8	41.9	42.1	48.2	———
Gold Ontonotes Data with Coref Chains								
EC	30.6	42.1	39.7	46.4	48.3	48.1	**51.5**	22.3%
EC-FM	26.6	39.9	37.6	45.4	46.7	46.2	49.8	———

the same test sets as the perplexity evaluation. We use mean reciprocal rank (MRR) and recall at 30 (Recall@30) to evaluate.

Results are provided in Table 4. Consistent with the results in the perplexity evaluation, LB outperforms other methods for both frame-chain and entity-centered SemLMs across all test sets. It is interesting to see that UNI performs better than BG in this prediction task. This finding is also reflected in the results reported in Rudinger et al. (2015). Though CBOW and SG have similar perplexity results, SG appears to be stronger in the narrative cloze test. With respect to the strongest baseline (UNI), LB achieves close to 20% relative improvement for Recall@30 metric on NYT hold-out data. On gold data, the frame-chain SemLMs get a relative improvement of 36.5% for Recall@30 while entity-centered SemLMs get 22.3%. For MRR metric, the relative improvement is around half that of the Recall@30 metric.

In the narrative cloze test, we also carry out an ablation study to remove the "FrameNet Mapping" step in Sec. 5.2 ("FC-FM" and "EC-FM" rows). The decrease in MRR and Recall@30 metrics further strengthens the argument that "FrameNet Mapping" is important for language modeling as it improves the generalization on frames.

We cannot directly compare with other related works (Rudinger et al., 2015; Pichotta and Mooney, 2016) because of the differences in data and evaluation metrics. Rudinger et al. (2015) also use the NYT portion of the Gigaword corpus, but with Concrete annotations; Pichotta and Mooney (2016) use the English Wikipedia as their data, and Stanford NLP tools for pre-processing while we use the Illinois NLP tools. Consequently, the eventual chain statistics are different, which leads to different test instances.[5] We counter this difficulty

[5]Rudinger et al. (2015) is similar to our entity-centered SemLM without discourse information. So, in Table 4, we

Table 5: **Co-reference resolution results with entity-centered SemLM features.** "EC" stands for the entity-centered SemLM. "TRI" is the tri-gram model while "LB" is the log-bilinear model. "p_c" means conditional probability features and "em" represents frame embedding features. "w/o DIS" indicates the ablation study by removing all discourse makers for SemLMs. We conduct the experiments by adding SemLM features into the base system. We outperform the state-of-art system (Wiseman et al., 2015), which reports the best results on CoNLL12 dataset. The improvement achieved by "EC_LB ($p_c + em$)" over the base system is statistically significant.

	ACE04	CoNLL12
Wiseman et al. (2015)	—	63.39
Base (Peng et al., 2015a)	71.20	63.03
Base+EC-TRI (p_c)	71.31	63.14
Base+EC-TRI w/o DIS	71.08	62.99
Base+EC-LB (p_c)	71.71	63.42
Base+EC-LB ($p_c + em$)	**71.79**	**63.46**
Base+EC-LB w/o DIS	71.12	63.00

by reporting results on "Gold PropBank Data" and "Gold Ontonotes Data". We hope that these two gold annotation datasets can become standard test sets. Rudinger et al. (2015) does share a common evaluation metric with us: MRR. If we ignore the data difference and make a rough comparison, we find that the absolute values of our results are better while Rudinger et al. (2015) have higher relative improvement ("Rel-Impr" in Table 4). This means that 1) the discourse information is very likely to help better model semantics 2) the discourse information may boost the baseline (UNI) more than it does for the LB model.

6.2 Evaluation of SemLM Applications

6.2.1 Co-reference Resolution

Co-reference resolution is the task of identifying mentions that refer to the same entity. To help improve its performance, we incorporate SemLM information as features into an existing co-reference resolution system. We choose the state-of-art Illinois Co-reference Resolution system (Peng et al., 2015a) as our base system. It employs a supervised joint mention detection and co-reference framework. We add additional features into the mention-pair feature set.

Given a pair of mentions (m_1, m_2) where m_1 appears before m_2, we first extract the corresponding semantic frame and the argument role label of each mention. We do this by following the procedures in Sec. 5. Thus, we can get a pair of semantic frames with argument information $(\text{fa}_1, \text{fa}_2)$. We may also get an additional discourse marker between these two frames, e.g. $(\text{fa}_1, \text{dis}, \text{fa}_2)$. Now, we add the following conditional probability as the feature from SemLMs:

$$p_c = p(\text{fa}_2|\text{fa}_1, \text{dis}).$$

We also add p_c^2, $\sqrt{p_c}$ and $1/p_c$ as features. To get the value of p_c, we follow the definitions in Sec. 4, and we only use the entity-centered SemLM here as its vocabulary covers frames with argument labels. For the neural language model implementations (CBOW, SG and LB), we also include frame embeddings as additional features.

We evaluate the effect of the added SemLM features on two co-reference benchmark datasets: ACE04 (NIST, 2004) and CoNLL12 (Pradhan et al., 2012). We use the standard split of 268 training documents, 68 development documents, and 106 testing documents for ACE04 data (Culotta et al., 2007; Bengtson and Roth, 2008). For CoNLL12 data, we follow the train and test document split from CoNLL-2012 Shared Task. We report CoNLL AVG for results (average of MUC, B^3, and $CEAF_e$ metrics), using the v7.0 scorer provided by the CoNLL-2012 Shared Task.

Co-reference resolution results with entity-centered SemLM features are shown in Table 5. Tri-grams with conditional probability features improve the performance by a small margin, while the log-bilinear model achieves a 0.4-0.5 F1 points improvement. By employing log-bilinear model embeddings, we further improve the numbers and we outperform the best reported results on the CoNLL12 dataset (Wiseman et al., 2015).

In addition, we carry out ablation studies to remove all discourse makers during the language modeling process. We re-train our models and study their effects on the generated features. Table 5 ("w/o DIS" rows) shows that without discourse information, the SemLM features would hurt the overall performance, thus proving the necessity of considering discourse for semantic language models.

6.2.2 Shallow Discourse Parsing

Shallow discourse parsing is the task of identifying explicit and implicit discourse connectives,

make a rough comparison between them.

Table 6: **Shallow discourse parsing results with frame-chain SemLM features.** "FC" stands for the frame-chain SemLM. "TRI" is the tri-gram model while "LB" is the log-bilinear model. "p_c", "em" are conditional probability and frame embedding features, resp. "w/o DIS" indicates the case where we remove all discourse makers for SemLMs. We do the experiments by adding SemLM features to the base system. The improvement achieved by "FC-LB ($p_c + em$)" over the baseline is statistically significant.

	CoNLL16 Test			CoNLL16 Blind		
	Explicit	Implicit	Overall	Explicit	Implicit	Overall
Base (Song et al., 2015)	89.8	35.6	60.4	75.8	31.9	52.3
Base + FC-TRI (q_c)	90.3	35.8	60.7	76.4	32.5	52.9
Base + FC-TRI w/o DIS	89.2	35.3	60.0	75.5	31.6	52.0
Base + FC-LB (q_c)	90.9	36.2	61.3	76.8	32.9	53.4
Base + FC-LB ($q_c + em$)	**91.1**	**36.3**	**61.4**	**77.3**	**33.2**	**53.8**
Base + FC-LB w/o DIS	90.1	35.7	60.6	76.9	33.0	53.5

determine their senses and their discourse arguments. In order to show that SemLM can help improve shallow discourse parsing, we evaluate on identifying the correct sense of discourse connectives (both explicit and implicit ones).

We choose Song et al. (2015), which uses a supervised pipeline approach, as our base system. The system extracts context features for potential discourse connectives and applies the discourse connective sense classifier. Consider an explicit connective "dis"; we extract the semantic frames that are closest to it (left and right), resulting in the sequence [f_1, dis, f_2] by following the procedures described in Sec. 5. We then add the following conditional probabilities as features. Compute

$$q_c = p(\mathrm{dis}|\mathrm{f}_1, \mathrm{f}_2).$$

and, similar to what we do for co-reference resolution, we add $q_c, q_c^2, \sqrt{q_c}, 1/q_c$ as conditional probability features, which can be computed following the definitions in Sec. 4. We also include frame embeddings as additional features. We only use frame-chain SemLMs here.

We evaluate on CoNLL16 (Xue et al., 2015) test and blind sets, following the train and development document split from the Shared Task, and report F1 using the official shared task scorer.

Table 6 shows the results for shallow discourse parsing with SemLM features. Tri-gram with conditional probability features improve the performance for both explicit and implicit connective sense classifiers. Log-bilinear model with conditional probability features achieves even better results, and frame embeddings further improve the numbers. SemLMs improve relatively more on explicit connectives than on implicit ones.

We also show an ablation study in the same setting as we did for co-reference, i.e. removing discourse information ("w/o DIS" rows). While our LB model can still exhibit improvement over the base system, its performance is lower than the proposed discourse driven version, which means that discourse information improves the expressiveness of semantic language models.

7 Conclusion

The paper builds two types of discourse driven semantic language models with four different language model implementations that make use of neural embeddings for semantic frames. We use perplexity and a narrative cloze test to prove that the proposed SemLMs have a good level of abstraction and are of high quality, and then apply them successfully to the two challenging tasks of co-reference resolution and shallow discourse parsing, exhibiting improvements over state-of-the-art systems. In future work, we plan to apply SemLMs to other semantic related NLP tasks e.g. machine translation and question answering.

Acknowledgments

The authors would like to thank Christos Christodoulopoulos and Eric Horn for comments that helped to improve this work. This work is supported by Contract HR0011-15-2-0025 with the US Defense Advanced Research Projects Agency (DARPA). Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. This material is also based upon work supported by the U.S. Department of Homeland Security under Award Number 2009-ST-061-CCI002-07.

References

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The berkeley framenet project. In *COLING/ACL*, pages 86–90.

N. Balasubramanian, S. Soderland, Mausam, and O. Etzioni. 2013. Generating coherent event schemas at scale. In *EMNLP*, pages 1721–1731.

D. Bamman and N. A. Smith. 2014. Unsupervised discovery of biographical structure from text. *TACL*, 2:363–376.

C. A. Bejan. 2008. Unsupervised discovery of event scenarios from texts. In *FLAIRS Conference*, pages 124–129.

E. Bengtson and D. Roth. 2008. Understanding the value of features for coreference resolution. In *EMNLP*.

N. Chambers and D. Jurafsky. 2008a. Jointly combining implicit constraints improves temporal ordering. In *EMNLP*.

N. Chambers and D. Jurafsky. 2008b. Unsupervised learning of narrative event chains. In *ACL*, volume 94305, pages 789–797.

N. Chambers and D. Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *ACL*, volume 2, pages 602–610.

N. Chambers. 2013. Event schema induction with a probabilistic entity-driven model. In *EMNLP*, volume 13, pages 1797–1807.

J. C. K. Cheung, H. Poon, and L. Vanderwende. 2013. Probabilistic frame induction. *arXiv:1302.4813*.

A. Culotta, M. Wick, R. Hall, and A. McCallum. 2007. First-order probabilistic models for coreference resolution. In *NAACL*.

L. Frermann, I. Titov, and Pinkal. M. 2014. A hierarchical bayesian model for unsupervised induction of script knowledge. In *EACL*.

M. Granroth-Wilding, S. Clark, M. T. Llano, R. Hepworth, S. Colton, J. Gow, J. Charnley, N. Lavrač, M. Žnidaršič, and M. Perovšek. 2015. What happens next? event prediction using a compositional neural network model.

M. Gutmann and A. Hyvarinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of HLT/NAACL*.

J. Irwin, M. Komachi, and Y. Matsumoto. 2011. Narrative schema as world knowledge for coreference resolution. In *CoNLL Shared Task*, pages 86–92.

B. Jans, S. Bethard, I. Vulić, and M. F. Moens. 2012. Skip n-grams and ranking functions for predicting script events. In *EACL*, pages 336–344.

P. Kingsbury and M. Palmer. 2002. From Treebank to PropBank. In *Proceedings of LREC-2002*.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *ICASSP*.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *NAACL*.

A. Mnih and G. Hinton. 2007. Three new graphical models for statistical language modelling. In *ICML*, pages 641–648.

A. Modi and I. Titov. 2014a. Inducing neural models of script knowledge. In *CoNLL*.

A. Modi and I. Titov. 2014b. Learning semantic script knowledge with event embeddings. In *ICLR Workshop*.

R. Mooney and G. DeJong. 1985. Learning schemata for natural language processing.

K.-H. Nguyen, X. Tannier, O. Ferret, and R. Besançon. 2015. Generative event schema induction with entity disambiguation. In *ACL*.

US NIST. 2004. The ace evaluation plan. *US National Institute for Standards and Technology (NIST)*.

B. Paul, B. Phil, and H. Hieu. 2014. Oxlm: A neural language modelling framework for machine translation. *The Prague Bulletin of Mathematical Linguistics*, 102(1):81–92.

H. Peng, K. Chang, and D. Roth. 2015a. A joint framework for coreference resolution and mention head detection. In *CoNLL*.

H. Peng, D. Khashabi, and D. Roth. 2015b. Solving hard coreference problems. In *NAACL*.

K. Pichotta and R. J. Mooney. 2014. Statistical script learning with multi-argument events. In *EACL*, volume 14, pages 220–229.

K. Pichotta and R. J. Mooney. 2016. Learning statistical scripts with lstm recurrent neural networks. In *AAAI*.

S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *CoNLL*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

V. Punyakanok, D. Roth, W. Yih, and D. Zimak. 2004. Semantic role labeling via integer linear programming inference. In *COLING*.

A. Rahman and V. Ng. 2011. Coreference resolution with world knowledge. In *ACL*.

D. Roth and D. Zelenko. 1998. Part of speech tagging using a network of linear separators. In *COLING-ACL*.

R. Rudinger, P. Rastogi, F. Ferraro, and B. Van Durme. 2015. Script induction as language modeling. In *EMNLP*.

R. C. Schank and R. P. Abelson. 1977. Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. In *JMZ*.

K. K. Schuler. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon.

Y. Song, H. Peng, P. Kordjamshidi, M. Sammons, and D. Roth. 2015. Improving a pipeline architecture for shallow discourse parsing. In *CoNLL Shared Task*.

A. Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*, volume 2002, page 2002.

I. Titov and E. Khoddam. 2015. Unsupervised induction of semantic roles within a reconstruction-error minimization framework. In *NAACL*.

Ben Wellner and James Pustejovsky. 2007. Automatically identifying the arguments of discourse connectives. In *Proceedings of the 2007 Joint Conference of EMNLP-CoNLL*.

T. Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3(1):1–191.

S. Wiseman, A. M. Rush, S. M. Shieber, and J. Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *ACL*.

N. Xue, H. T. Ng, S. Pradhan, R. P. C. Bryant, and A. T. Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In *CoNLL*.