Word Sense Disambiguation using localized RNNs

CS498A - UGP Presentation

Rushab Munot (14405)
CSE, IITK

Supervisor
Prof. Harish Karnick
CSE, IITK

November 20, 2017

Word Sense Disambiguation

- WSD addresses the task of classifying the sense of a word in use depending on its usage.

Example

1. It was in my interest to do so.
2. The bank's new policy boosted the interest rates on deposits.
3. The play was interesting.

Word Sense Disambiguation

- WSD addresses the task of classifying the sense of a word in use depending on its usage.

Example

1. It was in my interest to do so.
2. The bank's new policy boosted the interest rates on deposits.
3. The play was interesting.

- A more subtle example - *public-interest vs self-interest*

Example

1. In the people's interest - *public-interest*
2. My interest in the field - *self-interest*

Word Sense Disambiguation

- WSD addresses the task of classifying the sense of a word in use depending on its usage.

Example

1. It was in my interest to do so.
2. The bank's new policy boosted the interest rates on deposits.
3. The play was interesting.

- A more subtle example - *public-interest vs self-interest*

Example

1. In the people's interest - *public-interest*
2. My interest in the field - *self-interest*

- The context of a word is its immediate neighbourhood. In our approach, we model contexts using LSTM networks.

Wordnet Senses

- To disambiguate a word we do not need to know the complete meaning of the senses
- Only a certain number of characteristics are required

Wordnet Senses

- To disambiguate a word we do not need to know the complete meaning of the senses
- Only a certain number of characteristics are required
- For example, to disambiguate the word *set* in the following examples

Example

1. A *set* of rules - Noun
2. To *set* the volume - Verb

To distinguish between these two senses, we only need to know the POS tags

Wordnet Senses

- To disambiguate a word we do not need to know the complete meaning of the senses
- Only a certain number of characteristics are required
- For example, to disambiguate the word *set* in the following examples

Example

1. A *set* of rules - Noun
2. To *set* the volume - Verb

To distinguish between these two senses, we only need to know the POS tags

- However, to disambiguate between two verb senses of the same word, we need more informative tags

Example

1. To *set* fire - Verb $\Rightarrow$ To *start* a fire
2. To *set* the volume - Verb$\Rightarrow$ To *adjust* the volume

These two senses *cannot* be distinguished only with POS tags

WordNet Senses

- **WordNet Sense ⇒ sense_key = lemma % lex_sense**
  lex_sense = ss_type:lex_filenum:lex_id:head_word:head_id

Reference: https://wordnet.princeton.edu/man/senseidx.5WN.html

WordNet Senses

- **WordNet Sense ⇒ sense_key = lemma % lex_sense**
  lex_sense = ss_type:lex_filenum:lex_id:head_word:head_id
- **ss_type** is a 1-digit number between 1 and 5.
  It denotes the Synset type which is one of *Noun*, *Verb*, *Adjective*, *Adverb*, *Adjective Satellite* respectively

Reference: https://wordnet.princeton.edu/man/senseidx.5WN.html

WordNet Senses

- **WordNet Sense ⇒ sense_key = lemma % lex_sense**
  lex_sense = ss_type:lex_filenum:lex_id:head_word:head_id
- **ss_type** is a 1-digit number between 1 and 5.
  It denotes the Synset type which is one of *Noun, Verb, Adjective, Adverb, Adjective Satellite* respectively
- **lex_filenum** is the file number of the WordNet file (00 to 44) corresponding to the sense.
  - Based on syntactic category and logical groupings
  - For example, file 04 represents nouns denoting acts or actions;
    file 39 denotes perception related verbs (seeing, hearing, feeling)

Reference: https://wordnet.princeton.edu/man/senseidx.5WN.html

WordNet Senses

- **WordNet Sense ⇒ sense_key = lemma % lex_sense**
  lex_sense = ss_type:lex_filenum:lex_id:head_word:head_id
- **ss_type** is a 1-digit number between 1 and 5.
  It denotes the Synset type which is one of *Noun, Verb, Adjective, Adverb, Adjective Satellite* respectively
- **lex_filenum** is the file number of the WordNet file (00 to 44) corresponding to the sense.
  - Based on syntactic category and logical groupings
  - For example, file 04 represents nouns denoting acts or actions;
    file 39 denotes perception related verbs (seeing, hearing, feeling)
- **lex_id** along with the lemma differentiate between senses within a lexicographer file

Reference: https://wordnet.princeton.edu/man/senseidx.5WN.html

WordNet Senses

- **WordNet Sense ⇒ sense_key = lemma % lex_sense**
  lex_sense = ss_type:lex_filenum:lex_id:head_word:head_id
- **ss_type** is a 1-digit number between 1 and 5.
  It denotes the Synset type which is one of *Noun, Verb, Adjective, Adverb, Adjective Satellite* respectively
- **lex_filenum** is the file number of the WordNet file (00 to 44) corresponding to the sense.
  - Based on syntactic category and logical groupings
  - For example, file 04 represents nouns denoting acts or actions;
    file 39 denotes perception related verbs (seeing, hearing, feeling)
- **lex_id** along with the lemma differentiate between senses within a lexicographer file
- **head_word** is present only if ss_type=5.
  Satellite adjectives are basic adjectives like *dry* which when appended with a context enhance the meaning.
  "arid" = "dry" + "climate"; "thirsty" = "dry" + "throat"; etc
- **head_id** is similar to lex_id

Reference: https://wordnet.princeton.edu/man/senseidx.5WN.html

The model

- The model is a (backward directional) LSTM network (with attention) for generating context embeddings
- The model is localized at the word level i.e. one LSTM network per word

The model

- The model is a (backward directional) LSTM network (with attention) for generating context embeddings
- The model is localized at the word level i.e. one LSTM network per word
- The attention mechanism is word dependent but context independent i.e. fixed for every word.

The model

- The model is a (backward directional) LSTM network (with attention) for generating context embeddings
- The model is localized at the word level i.e. one LSTM network per word
- The attention mechanism is word dependent but context independent i.e. fixed for every word.
- The context windows have a fixed size (may vary for different words). In our experiments the optimum size was observed to be 10.

The model

- The model is a (backward directional) LSTM network (with attention) for generating context embeddings
- The model is localized at the word level i.e. one LSTM network per word
- The attention mechanism is word dependent but context independent i.e. fixed for every word.
- The context windows have a fixed size (may vary for different words). In our experiments the optimum size was observed to be 10.
- We observed that self attention mechanisms tend to have a fixed profile, with little deviations depending on the contexts.

The Model

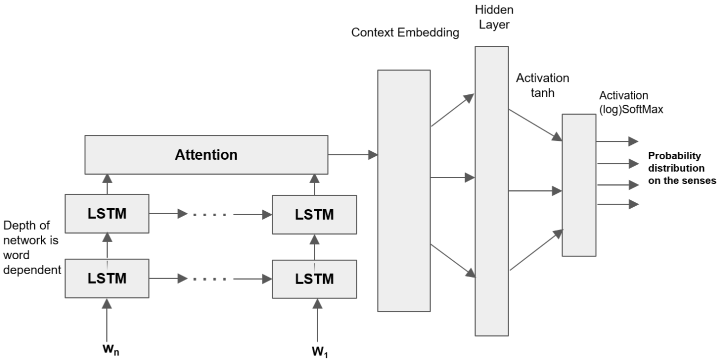

Figure: The Model and the Attention Mechanism

The Model

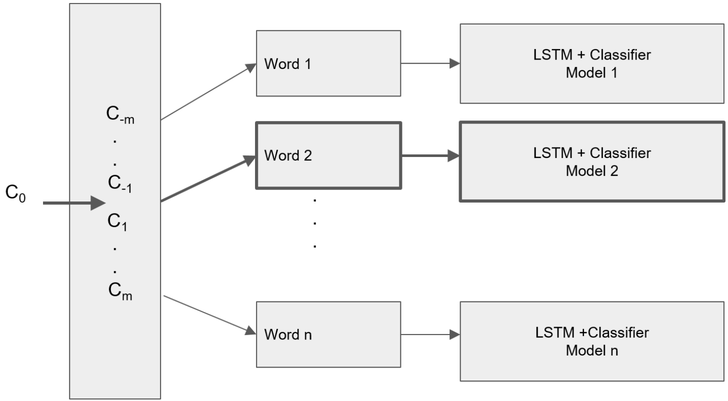

Figure: One model per word

Image taken from my UGP titled WSD using RNNs for context embedding (CS396A, II Semester, 2016-17)

Generating Data

- For the OMSTI dataset, more data is generated by gathering all examples have the same sense as the given sense.
- This helps to have more data with varied sentences.

Generating Data

- For the OMSTI dataset, more data is generated by gathering all examples have the same sense as the given sense.
- This helps to have more data with varied sentences.
- This approach does not always work with the Lexical Sample Task where the data is scarce (200 sentences per word with many words having 6-7 senses). The data is also highly unbalanced.

Datasets

- Senseval-2 dataset of 4 words - hard, line, interest, serve
  About 4000 sentences for hard, line, serve; 2000 for interest
  Only those senses with more than 300 senses

Datasets

- Senseval-2 dataset of 4 words - hard, line, interest, serve
  About 4000 sentences for hard, line, serve; 2000 for interest
  Only those senses with more than 300 senses
- Certain words from the One-million word corpus interest $=$ 4 senses (500 each), position $=$ 5 senses (500 each), serve $=$ 8 senses (6 senses 500, remaining 2 have lesser instances)

Datasets

- Senseval-2 dataset of 4 words - hard, line, interest, serve
  About 4000 sentences for hard, line, serve; 2000 for interest
  Only those senses with more than 300 senses
- Certain words from the One-million word corpus interest = 4 senses (500 each), position = 5 senses (500 each), serve = 8 senses (6 senses 500, remaining 2 have lesser instances)
- Senseval-2 English Lexical Sample Task

Accuracy

Word	#Senses	BLSTM Embedding		Attention Mechanism	
		Accuracy	**F1**	**Accuracy**	**F1**
hard	3	90.90	79.28	91.55	79.99
interest	4	87.67	80.42	91.85	84.11
serve	4	84.79	81.76	86.8	83.705
line	6	79.05	70.65	83.5	77.21

Table: Accuracy and F1(macro) score on the Senseval-2 dataset

- Accuracy not a good measure, as data is unbiased (for Senseval-2), F1 is a better measure

Accuracies - Senseval-2, English Lexical Sample Task

Word	#Instances	#Senses	Score*
art	196	8	60.2
authority	184	9	62.5
bar	304	13	51.7
blind	108	9	80.0*
bum	92	6	75.6
chair	138	4	81.7*
channel	145	9	41.7
child	129	8	59.4
church	128	5	62.5
colorless	67	3	60.0*
cool	106	7	50.0
day	289	9	71.0

Table: Accuracy (as evaluated by their scorer)* on the Senseval-2, English Lexical Sample Task

Here we have predicted only one sense (instead of multiple senses). We have precision = recall and fine grained = coarse grained)

CAUSE - OMSTI - Accuracy Plots

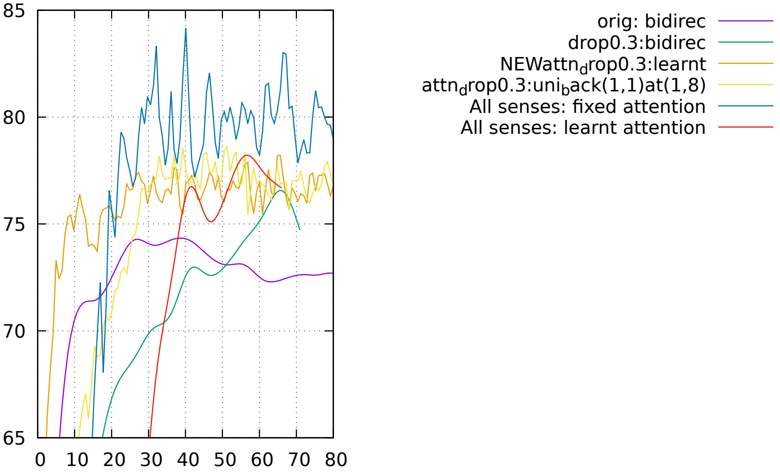

Figure: *cause*: OMSTI dataset - Comparison between different methods

SERVE- OMSTI - Accuracy Plots

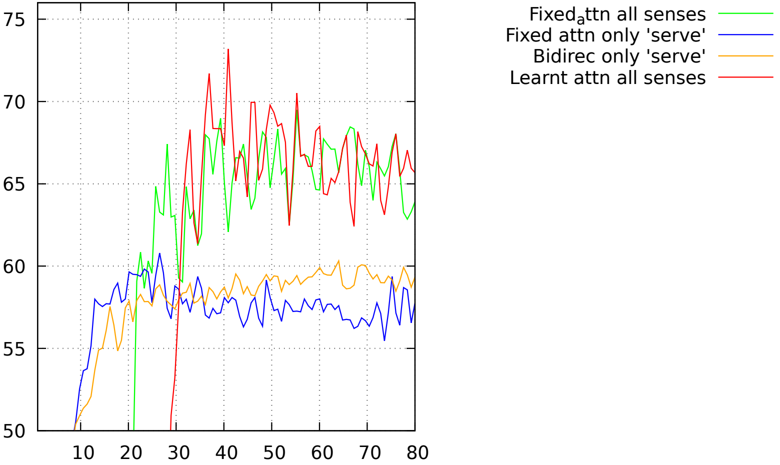

Figure: *serve*: OMSTI dataset - Comparison between different methods

HARD: 4 Words Dataset - Accuracy Plots

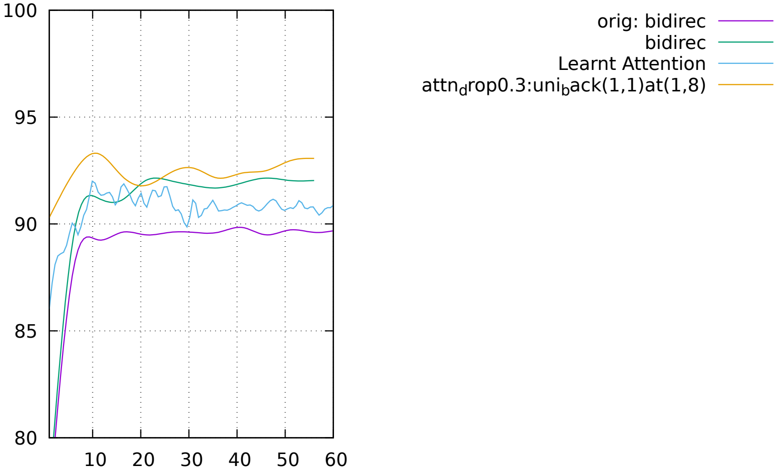

Figure: *hard*: Four word dataset - Comparison between different methods

SERVE- 4 Words Dataset - Accuracy Plots

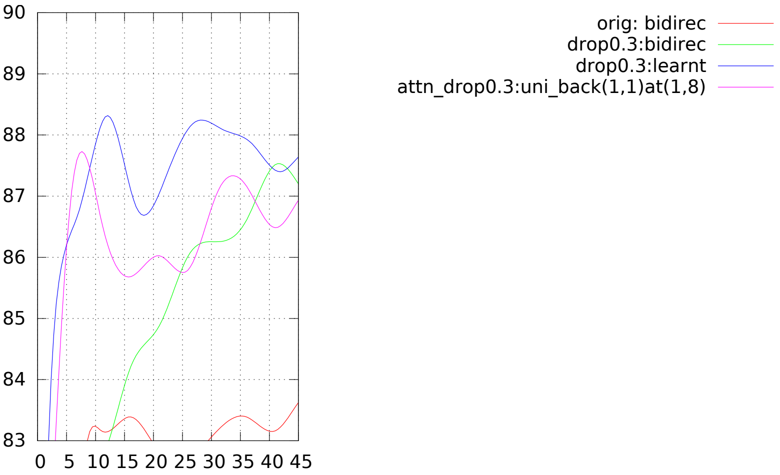

Figure: *serve*: Four word dataset - Comparison between different methods

Future Work

- The model can be used for sense vector generation. One such approach would be to modify the WSD layer in - *Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space'* (Neelakantan et Al., 2014).
- To develop a model that inherently uses the hierarchical nature of senses.

Acknowledgements

- Datasets from Senseval-2 and One million word corpus (Kaveh Taghipour and Hwee Tou Ng, 2015 - One Million Sense-Tagged Instances for Word SenseDisambiguation and Induction)
- Word Sense Disambiguation: A Survey - R. Navigli, 200
- WordNet 3.1 - https://wordnet.princeton.edu/

THANK YOU