NLP-1 Pen&Paper 3 - ard snijders - 12854913

(a) - Why do compositional semantics struggle with these phrases?

(i) it snows

In this sentence, *it* is a pleonastic pronoun, which means it is a semantically empty pronoun. these pronouns do not refer to anything; a system would thus have to find a way to learn that these instances are pleonastic and should not be counted as mentions in coreference tasks.

(ii) rock the boat

Not all phrases are interpreted compositionally; *rock the boat* is an idiomatic expression which means something along the lines of "say or do something to disturb an existing situation and upset people." Simultaneously, these phrases **can** be compositionally interpreted (in theory, a boat can be rocked), so we cannot simply block them. Thus, a system would have to learn a way to deal with this.

(iii) enjoy a cigarette
Enjoy a book

These phrases are examples of elliptical construction where meaning arises through composition; specifically, logical metonymy; the combination of an event-subcategorizing verb (*enjoy*) with entity-denoting objects (*cigarette, book)*, such that the interpretation of the verb phrase requires the retrieval of non-explicitized or *covert* events (Chersoni et al, 2017) - in this case; the acts of smoking and reading, respectively. Possible, a model would have to learn to obtain the appropriate covert event to correctly interpret the verb.

(iv) inject life
Inject hydrogen

These phrases are examples of meaning transfer and additional connotations that arise through composition, specifically metaphors in this instance; here, the word *inject* takes on a different meaning depending on the context in which it is used. Thus, the model would have to learn how to distinguish between these different uses of *inject*.

b) [3 points]

(i) Vector mixture models are used to model sentence composition in a distributional space. Three disadvantages of vector mixture models:

- Vector mixture models are commutative and therefore do not take word order into account: you essentially get a bag of words, and re-arranging words will not affect the outcome of the resulting sentence representation. Vector mixture models can thus capture topical information but are less suited for capturing more fine-grained syntactic / structural information
- Vector mixture models are less suitable to model function words; we can typically learn good representations for nouns and adjectives, but less so for conjunctions or prepositions, because those occur with many other words (other contexts) and their distributions won't be as specific compared to e.g. nouns.
- Vector mixture models are insensitive to sentence semantics and cannot capture certain linguistic phenomena, e.g. irony/sarcasm/humor, or different word senses.

(ii) How lexical function models address these limitations:

Lexical function models address some of the limitations of vector mixture models. For instance, they address the disadvantage of mixture models being less suitable for modelling function words, by making a distinction between words whose meaning is directly determined by their distributional behaviour, and words that act as functions, transforming the distributional profile of other words, e.g. verbs and adjectives. If trained properly, lexical function models can also deal with polysemy in use of e.g. metaphors. However, the lexical function model does not provide a complete solution to the bag-of-words limitation of mixture models - in that regard, one can debate the validity of such a comparison; vector mixture models can be used for entire sentence representations, whereas (to my knowledge) lexical function models can only be applied to e.g. adjective-noun phrases, and in that respect do not solve that problem.

(c) [3 points]
Adjectives can be represented as lexical functions through the use of a parameter matrix.

For instance, consider the noun *dog* and adjective *old*. The lexical function is then modelled as *old dog* = *old (dog)*. A parameter matrix A_{old} can then serve as a function for adjective *old*, and this matrix can be used to linearly transform the noun vector n_{dog}, such that we get a vector for *old dog*.

In order to learn a set of parameters that allow for prediction of vectors of adjective-noun phrases, we first obtain a distributional vector n_j for each noun in the lexicon. We then collect frequent adjective-noun pairs (a_j, n_j), arguably by use of some threshold to discard infrequent pairings. Lastly, we obtain the distributional pair vector p_{ij} through use of a conventional distributional semantic model. We can now learn the parameters of an adjective matrix A with linear regression, by minimizing the squared error loss L:

$$L(\mathbf{A}_i) = \sum_{j \in \mathcal{D}(a_i)} \|\mathbf{p}_{ij} - \mathbf{A}_i \mathbf{n}_j\|^2$$

In other words, adjective matrix A_i is learned until it can serve as an appropriate lexical function for noun n_j, i.e. the transformation of a noun n_j by performing $A_i n_j$ should result in a vector that approximates the 'true' associated pair vector p_{ij}.

(d) The above lexical function model of adjective meaning can be used to:

(i) paraphrase adjective-noun phrases, obtaining paraphrase pairs such as *smart students → clever students*. [2 points]

Assuming that adjective matrices for the current adjective (*smart)* and desired adjective (*clever)* are known, it should be possible to obtain the distributional vector of *students* by reversing the linear transformation (multiplying the vector *smart students* with the inverse matrix of *smart*), after which we can multiply the vector of *students* with the adjective matrix for *clever*, and this should result in a vector close to the true vector of *clever students*.

(ii) disambiguate the meaning of the adjective, for instance the meaning of *warm* in *warm tea vs. warm person*, by ……..

I am not sure about this question; what I gathered from the lecture, in principle, we use a single representation for all different word senses, and it is assumed that ambiguity can be dealt with by relying on context - a few exceptions are given where researchers handle this problem with prior disambiguation, but in general, a single representation is used. So in that respect, I'm a bit puzzled on how the lexical function modelling of adjective meaning could be used to deal with this sort of polysemy, since - in the **above model** there is a single adjective matrix for *warm* which is used for all word senses.

That being said, it could be possible to deal with adjective ambiguity by taking into account the distribution of the phrase it occurs in, after which the adjective could be classified as the word sense appropriate for that context

As proposed by Gutierrez et al (2016), one could learn separate adjective matrices for all literal / metaphorical word senses of an adjective (though at the expense of having less data for each of the distinct word senses to train on) - then, for an unseen adjective-noun phrase (involving the adjective for which different matrices are already learnt), one could calculate cosine similarities to determine whether that unseen phrase is most similar to a literal or metaphorical adjective matrix, and classify the unseen adjective as literal/metaphorical accordingly, as formulated in the following expression:

$$\cos(\mathbf{p}_i, \hat{\mathbf{A}}_{\mathrm{MET}(a)}\mathbf{n}_i) < \cos(\hat{\mathbf{A}}_{\mathrm{LIT}(a)}\mathbf{n}_i).$$

Where cos denotes cosine similarity, p_i denotes an unseen phrase, $\widehat{A}$ the estimated matrix for adjective a, and the MET/LIT subscripts denoting the metaphorical and literal word senses, respectively.

(e) The two sentences in each sentence pair below are linked by a particular rhetorical relation. Which rhetorical relation does each sentence pair exhibit?

> *(i) The use of diesel in transport has come under increasing scrutiny in recent years. According to WHO, around three million deaths every year are linked to exposure to outdoor air pollution.*
>
> Explanation - the use of diesel in transport has come under scrutiny *because* three million deaths every year are linked to exposure to outdoor air pollution. The first sentence is the nucleus, the second sentence is the satellite.
>
> *(ii) Nitrogen oxides can help form ground level ozone. This can exacerbate breathing difficulties.*
>
> Elaboration - the satellite gives additional information about the nucleus. The first sentence is the nucleus, the second sentence is the satellite.
>
> *(iii) Paris has already taken a series of steps to cut the impact of diesel cars and trucks. Vehicles registered before 1997 have already been banned from entering the city.*
>
> I would argue this sentence exhibits both elaboration and evidence relations.
>
> - in the case of elaboration, the second sentence is the satellite and provides more information about the first sentence/nucleus.
> - In the case of justification/evidence, the second sentence is again the satellite and provides evidence for the claim made in the first sentence/nucleus.

(note: I borrow the rhetorical relation terms from jurafsky & martin, who use slightly different terminology compared to the slides)

(f) Specify which salient rule is used in resolve the highlighted anaphora in the given discourse.

*(i) Lee bought his first car in 2005 and his second car, last year. He now drives **it** to work.*

Recency: *it* likely refers to *his second car*, because it's an entity more recently uttered than *his first car*.

*(ii) Lara scolded Maria for breaking the glass; **she** still couldn't contain her anger.*

Grammatical role: *She* likely refers to Lara because Subject > Object > everything else, and here Lara is the subject and Maria the object. Entities in the subject position are more salient than those in the object position, because entities are usually introduced as subjects.

*(iii) Elizabeth first danced with Mr Darcy. Anna danced with **him** next.*

Parallelism: Entities which share the same role as the pronoun in the same sort of sentence are preferred. *Him = Mr. Darcy* both share the same role.

Recency: entities introduced in recent utterances are more salient than those introduced from utterances further back; thus, *him* is likely to refer to *Mr Darcy* because it is the most recently introduced entity.

(g) Consider the following discourse:

```
Arya went to Hilda's car showroom to check out a Fiat Linea.
She decided to buy it, after hours of inspection.
```

(i)

pron	**ante**	**cat**	**num**	**gen**	**same**	**dist**	**role**	**par**	**form**
she	Arya	f	t	t	f	1	subj	t	prop
she	Hilda	f	t	t	f	1	obj	f	prop
it	car showroom	f	t	t	f	1	obj	t	definite
it	A Fiat Linea	f	t	t	f	1	obj	t	indefinite

(ii) Describe how this feature vector can be used to resolve the anaphora assuming a supervised approach. [1 point]

This feature vector could be used to resolve the anaphoras *she* and *it* by use of a classification algorithm, such as SVM or random forest regression. For instance, a random forest regressor might learn that the antecedent candidates in the subject role (e.g. Arya) tend to be more salient.

(iii) Mention two problems with this simple classification approach. [1 point]

Its difficult to implement the 'repeated mention' effect (entities that are mentioned often are preferred) by just looking at different pairs, and so we lose out on that information.

Another reason why this approach can be problematic is that typically text does not so much have pairs, but rather coreference chains; typically you talk about the same entity multiple times, and it's generally better to look at the whole chain (which is not possible with a model that uses isolated pairs).