A Quick Summary: Compositional Morphology for Word Representations and Language Modelling

Original Paper: http://proceedings.mlr.press/v32/botha14.pdf

9 March 2019

1 Ideas:

(a) Seek a compromise: Retain the unsupervised nature of CSLM (Continuous space language models) and incorporate a priori linguistic knowledge. Specifically, morphologically related words should share statistical strength in spite of differences in surface form

(b) The model introduced here is the Additive Log-Bilinear Model.

2 Explanations:

(a) Each word should be thought of as a sum of its constituent morphemes, eg:

$$\overrightarrow{\text{imperfection}} = \overrightarrow{in} + \overrightarrow{perfect} + \overrightarrow{ion}$$

The surface form of a word is also registered as a factor to account for noncompositional constructions and ordering of words (hangover $\neq$ overhang), so the above example becomes

$$\overrightarrow{\text{imperfection}} = \overrightarrow{in} + \overrightarrow{perfect} + \overrightarrow{ion} + \overrightarrow{imperfection}$$

(b) The (traditional) Log-Bilinear Language model is formulated in the following way:

The vector for the next word $\mathbf{p}$ is a function of the context vectors $\mathbf{q_j} \in \mathbb{R}^{\mathbf{d}}$ of the preceding words:

$$\mathbf{p} = \sum_{j=1}^{n-1} \mathbf{q}_j C_j$$

Where $C_j \in \mathbb{R}^{d \times d}$, $j = 1, ..., n-1$.

$v(w)$ indicates how well the observed word w fits the prediction $\mathbf{p}$ and is defined as $v(w) = \mathbf{p} \cdot \mathbf{r}_w + b_w$. Then, taking a softmax gives individual word probabilities as:

$$P(w_i|w_{i-n+1}^{i-1}) = \frac{exp(v(w_i)}{\sum_{v \in \mathcal{V}} exp(\mathcal{V}(v))}$$

Thus, the model is subsequently denoted as **LBL** with parameters $\Theta_{LBL} = (C_j, Q, R, \mathbf{b})$, where we have $Q, R \in \mathbb{R}^{|\mathcal{V}| \times d}$ and $\mathbf{b} \in \mathbb{R}^{|\mathcal{V}|}$

Now, we see how the additive Log-Bilinear Model differs.

i $\tilde{\mathbf{r}}$ and $\tilde{\mathbf{q}}_j$ are now the composed word vectors.

ii Representation matrices are now $Q^{(f)}, R^{(f)} \in \mathbb{R}^{|\mathcal{F}| \times d}$, so that we have some $M \in \mathbb{Z}_+^{\mathcal{V} \times |\mathcal{F}|}$ such that $R = MR^{(f)}$ and $Q = MQ^{(f)}$

iii There are two obvious variations of the LBL_{++}, which apply the factorisation on either the simple word vectors or the context word vectors.

3 Results:

(a) Better performance for CSLM and n-gram MKN models.

4 Notes:

(a) I think this is quite interesting. Although this seems to be language specific and thus lacking generality, it may be a key to further understanding language modeling in English.

(b) What if instead of predicting words we focus on predicting morphemes? This decreases the vocabulary, while not necessarily decreasing the representational capacity of the model.