Language Models Based on Semantic Composition

Jeff Mitchell and Mirella Lapata

Matija Hanževački, Saarland University

Recent Advances in Computational Semantics
Seminar

December 2013

Introduction

Introduction

- Proposed a novel statistical language model to capture long-range semantic dependencies.
- Applied to the problem of constructing predictive history representations of upcoming words.
- Effects measured as reductions in *perplexity*.

Language Models

- Assign probabilities to sequences of words.
- Estimated as the product of conditional probabilities of words given their history of preceding words:

$$P(w_i|h_i) \qquad h_i \equiv w_1^{i-1}$$

- In practice, history spans up to 3-5 words back.

Related Work

- Language models have compromised ability of capturing long-range dependencies (given local history).
- Handled in the literature using modulation of probabilities by dependencies which extend to words beyond the *n*-gram horizon.
- Various ways of capturing syntactic and semantic dependencies on a word level.
- Authors propose composing the meaning of history using their vector composition framework described in (Mitchell and Lapata, 2008).

Composition Models

Composition Framework

- Using the framework formulated as a function of two vectors.
- Addition lumps the contents of the vectors together.
- Multiplication picks out the content relevant to their combination by scaling each component of one with the strength of the corresponding component of the other vector.

$$\mathbf{h} = f(\mathbf{u}, \mathbf{v}) \quad h_i = u_i + v_i \quad h_i = v_i \cdot u_i$$

The Probabilistic Argument

- Define semantic vector components as:

$$v_i = \frac{p(context_i|target)}{p(context_i)}$$

- Multiplicative model represents distributional properties of the phrase *w1 and w2* and the additive model represents *w1 or w2*.

$$h_i = v_i \cdot u_i = \frac{p(c_i|w_1)}{p(c_i)} \frac{p(c_i|w_2)}{p(c_i)}$$

$$h_i = \frac{p(w_1|c_i)p(w_2|c_i)}{p(w_1)p(w_2)}$$

$$h_i \approx \frac{p(w_1 w_2|c_i)}{p(w_1 w_2)} = \frac{p(c_i|w_1 w_2)}{p(c_i)}$$

$$p(c_i|x) = \frac{1}{2}p(c_i|w_1) + \frac{1}{2}p(c_i|w_2)$$

$$x_i = \frac{1}{2}\frac{p(c_i|w_1)}{p(c_i)} + \frac{1}{2}\frac{p(c_i|w_2)}{p(c_i)}$$

Estimating Probabilities

- Semantic coherence commonly measured using the cosine measure:

$$sim(\mathbf{w},\mathbf{h}) = \frac{\mathbf{w}\cdot\mathbf{h}}{|\mathbf{w}||\mathbf{h}|} \qquad \mathbf{w}\cdot\mathbf{h} = \sum_i w_i h_i$$

- Using the former definition of vector components:

$$\mathbf{w}\cdot\mathbf{h} = \sum_i \frac{p(c_i|w)}{p(c_i)}\frac{p(c_i|h)}{p(c_i)}$$

$$p(w|h) = p(w)\sum_i \frac{p(c_i|w)}{p(c_i)}\frac{p(c_i|h)}{p(c_i)}p(c_i)$$

$$= \sum_i p(w|c_i)p(c_i|h)$$

$$\mathbf{h}_n = f(\mathbf{w}_n,\mathbf{h}_{n-1})$$

$$\mathbf{h}_1 = \mathbf{w}_1$$

$$h_i = \frac{\hat{h}_i}{\sum_j \hat{h}_j\cdot p(c_i)}$$

Integrating with Other Language Models

- Based on the previous language model:

$$p(w|h) = p(w) \cdot \Delta(w,h)$$

$$\Delta(w,h) = \sum_i \frac{p(c_i|w)}{p(c_i)} \frac{p(c_i|h)}{p(c_i)} p(c_i)$$

- Integrating the *n*-gram model:

$$\hat{p}(w_n) = p(w_n|w_{n-2}^{n-1}) \cdot \Delta(w_n, h)$$

$$p(w_n|w_{n-2}^{n-1}, h) = \frac{\hat{p}(w_n)}{\sum_w \hat{p}(w)}$$

Experiments

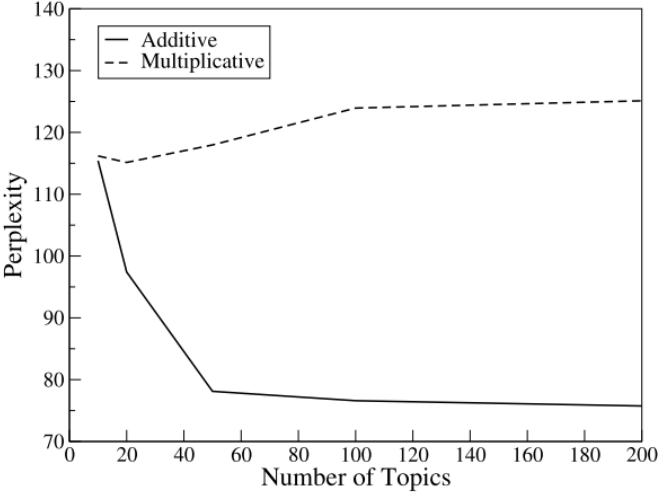
Experimental Setup

- Experimenting with additive and multiplicative composition functions, and with two semantic representations (LDA and the Simple Semantic Space Model).
- Models compared against a structured language model by (Roark 2001).
- Evaluated using *perplexity* – quantified degree of unpredictability in a probability distribution (lower is better).
- Trained on the BLLIP corpus (news texts) of about 38.5 million words.

Results

Model	Perplexity
n-gram	78.72
n-gram+Add$_{\text{SSM}}$	76.65
n-gram + Multiply$_{\text{SSM}}$	75.01
n-gram+Add$_{\text{LDA}}$	76.60
n-gram+Multiply$_{\text{LDA}}$	123.93
parser	173.35
n-gram + parser	75.22
n-gram + parser + Add$_{\text{SSM}}$	73.45
n-gram + parser + Multiply$_{\text{SSM}}$	71.32
n-gram + parser + Add$_{\text{LDA}}$	71.58
n-gram + parser + Multiply$_{\text{LDA}}$	87.93

Results Cont.



Conclusion

Conclusion

- Proposed the use of vector composition models for language modelling.
- Enhanced a trigram model with long-range semantic dependencies.
- Compared addition and multiplication based models and examined the influence of the underlying semantic space on the composition task.
- Best results with multiplicative composition function in a simple semantic space.

Thank you! Questions?