A step-by-step derivation of the equations in the paper *Embedding a semantic network in a word space*

Richard Johansson
richard.johansson@gu.se

1 General formulation

The general optimization problem stated in the paper by Johansson and Nieto Piña (2015) is the following:

$$\begin{aligned} \underset{E,p}{\text{minimize}} \quad & \sum_{i,j,k} w_{ijk} \Delta(E(s_{ij}), E(n_{ijk})) \\ \text{subject to} \quad & \sum_j p_{ij} E(s_{ij}) = F(l_i) \qquad \forall i \\ & \sum_j p_{ij} = 1 \qquad \forall i \\ & p_{ij} \geq 0 \qquad \forall i, j \end{aligned} \tag{1}$$

Explanation of the notation:

- The lexicon defines ambiguous *lemmas* l_i (e.g. *rock*), and *senses* $s_{i1}, \ldots, s_{im_i}$ (e.g. *rock-1* corresponding to the material and *rock-2* corresponding to the type of music).
- For each lemma l_i, there is a given *lemma embedding* $F(l_i)$, a D-dimensional vector typically computed by a distributional model such as `word2vec` or GloVe.
- For each sense s_{ij}, we create a *sense embedding* $E(s_{ij})$, again a D-dimensional vector.
- The lemma embeddings can be decomposed into a *mix* (e.g. a convex combination) of sense vectors, for instance $F(\textit{rock}) = 0.3 \cdot E(\textit{rock-1}) + 0.7 \cdot E(\textit{rock-2})$. The "mix variables" p_{ij} are non-negative and sum to 1 for each lemma.
- The intuition of the optimization that each sense s_{ij} should be "close" to a number of other concepts, called the *network neighbors*, that we know are related to it, as defined by a semantic network. For instance, *rock-2* might be defined by the network to be related to other types of music.
- Each network neighbor n_{ijk} is associated with a weight w_{ijk}. A higher weight means that we should work hard to force the sense to be close to this neighbor.
- The notion of "closeness" is formalized as a *distance function* defined on the D-dimensional vector space. In the rest of the paper, this will be assumed to be the squared Euclidean.
- The goal of the optimization problem is to find the sense embeddings and the mix variables. Everything else is assumed to be given to the algorithm as input.

2 Simplified optimization problem

Since the general problem (1) is hard to solve directly, we resort to an iterative approximation algorithm, where we consider one lemma l_i at a time, and solve the optimization problem just for this lemma:

$$L_i(E) = \sum_{jk} w_{ijk} \|E(s_{ij}) - E(n_{ijk})\|^2 \tag{2}$$

The next goal is to show that the full optimization problem (where we search for the optimal sense embeddings $E(s_{ij})$ and the mix variables p_{ij}) can be rewritten, so that the sense embeddings can be written in closed form if the p_{ij} are given. As mentioned in the paper, this means that we have reduced a problem of optimizing $m_i \cdot D + m_i$ variables to one where we just have $m_i - 1$ variables.

When the p_{ij} are given, the only constraint remaning is the one that forces the weighted sense embeddings to sum to the lemma embedding. To reduce the notational clutter, we drop the i index and abuse the notation a bit: we write just s_j when we mean the embedding $E(s_{ij})$ of the sense s_{ij} of lemma l_i, and similarly we write just l for the lemma embedding $F(l_i)$. After all the simplifications, we get a quadratic problem with equality constraints:

$$\begin{aligned} \underset{s}{\text{minimize}} \quad & \sum_{j,k} w_{jk} \|s_j - n_{jk}\|^2 \\ \text{subject to} \quad & \sum_j p_j s_j = l \end{aligned} \tag{3}$$

3 Deriving the closed-form solution

We introduce a Lagrangian λ into (3) for the equality constraint, and get a dual unconstrained optimization problem. Here, λ is a vector having the same dimensionality D as the embeddings. (The number 2 is arbitrary and was selected to simplify the equations later on.)

$$L'_i(s, \lambda) = \sum_{jk} w_{jk} \|s_j - n_{jk}\|^2 - 2\lambda(\sum_j p_j s_j - l) \tag{4}$$

We take the partial derivative of the dual (4) with respect to the m-th dimension of the sense embedding s_j, and set it to zero:

$$\frac{\partial}{\partial s_j^{(m)}} L'_i(s, \lambda) = 2 \sum_k w_{jk} (s_j^{(m)} - n_{jk}^{(m)}) - 2 p_j \lambda^{(m)} = 0 \tag{5}$$

As in the paper, we introduce the notion of the *weighted centroid* to simplify the notation.

$$c_j := \frac{\sum_k w_{jk} n_{jk}}{\sum_k w_{jk}} \tag{6}$$

We solve equation (5) for $s_j^{(m)}$, and use the notation of c_j to keep things compact:

$$s_j^{(m)} = c_j^{(m)} - \frac{p_j}{\sum_k w_{jk}} \lambda^{(m)} \tag{7}$$

Next, we take the partial derivative of the dual (4) with respect to the m-th dimension of the Lagrangian, and set it to zero. (This corresponds to the equality constraint.)

$$\frac{\partial}{\partial \lambda^{(m)}} L'_i(s, \lambda) = \sum_j p_j s_j^{(m)} - l^{(m)} = \sum_j p_j \left(c_j^{(m)} - \frac{p_j}{\sum_k w_{jk}} \lambda^{(m)} \right) - l^{(m)} = 0$$

We rearrange a bit...

$$\left(\sum_j p_j c_j^{(m)} - l^{(m)}\right) - \sum_j \frac{p_j^2}{\sum_k w_{jk}} \lambda^{(m)} = 0$$

...and solve for $\lambda^{(m)}$:

$$\lambda^{(m)} = \frac{1}{\sum_j \frac{p_j^2}{\sum_k w_{jk}}} \left(\sum_j p_j c_j^{(m)} - l^{(m)}\right)$$

Finally, as in equation (4) in the paper, we introduce the notion of the *residual*, again with the purpose of making the notation compact:

$$r := \frac{1}{\sum_j \frac{p_j^2}{\sum_k w_{jk}}} \left(\sum_j p_j c_j - l\right) \tag{8}$$

This leads us to the main result, corresponding to equation (5) in the paper.

$$s_j = c_j - \frac{p_j}{\sum_k w_{jk}} r \tag{9}$$

4 A few observations

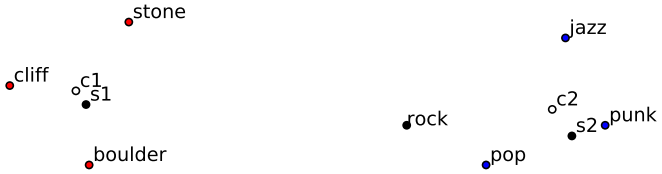
- If p_j is 0 for some sense s_j, then its sense embedding is equal to the weighted centroid c_j. This means that it is completely determined by its network neighbors, and uninfluenced by the lemma embedding. This follows trivially from (9). Intuitively, this means that this sense does not occur in the corpora used to compute the lemma embedding.
- Conversely, if p_j is 1 for some sense s_j, then it is equal to the lemma embedding l, and completely uninfluenced by its network neighbors: intuitively, this sense dominates the corpus. To see why this is the case, we first note that all other p-s are zero, so then we get

$$r = \sum_k w_{jk} \left(c_j - l\right) \qquad\qquad s_j = c_j - \frac{1}{\sum_k w_{jk}} r = l$$

- If we didn't have the equality constraint, the optimal s_j would be equal to the weighted centroid c_j. The residual is zero if the solution of the unconstrained problem already satisfies the constraints.

5 Example

The following figure shows a hypothetical example.



In this example, we assume the following input:

- an embedding for the lemma *rock*;
- *rock* has one sense whose network neighbors are *cliff*, *stone*, and *boulder*;

- *rock* has another sense whose network neighbors are *jazz*, *pop*, and *punk*;
- all neighborhood weights w_{jk} are equal to 1.

Since all the neighborhood weights are equal, the centroids c_1 and c_2 correspond to the points "in the middle" of the two neighborhoods, respectively. In this case, we set p_1 closer to 0 and p_2 closer to 1, so s_2 is closer to the lemma embedding. Note that s_1, s_2, and the lemma embedding form a straight line.

References

Richard Johansson and Luis Nieto Piña. 2015. Embedding a semantic network in a word space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1428–1433, Denver, United States.