Lecture 13 - Semantic Spaces

Summary

Up to this point when we've been considering BOW models, we've always made the assumption that words are independent. This is clearly not the case with natural language however as there are many words which for example are synonymous. In this lecture we'll look at techniques which begin to deal with this problem and allow better measurements of distances between both words and documents to be made. We'll also look at extreme examples of dealing with synonyms where the words are in different languages, and even different modalities.

Key points

Mining Distributional Semantics

- *Distributional Hypothesis:* words that are close in meaning will occur in similar pieces of text
  - Can we exploit this to uncover hidden meaning?
    - **Latent Semantics Analysis (LSA)**
    - Topic Modelling
- Latent Semantic Analysis is a technique that exploits distributional semantics to project words into a *term* space where their vectors are not orthogonal. It also produces a *document* space where the cosine similarity between documents is based on the non-orthogonal word vectors (and hence should be a better representation of true distance)
  - Start by considering a **term-document matrix** which described occurrences of terms in documents
    - Clearly going to be sparse
    - Could be weighted (c.f. TF-IDF)
  - LSA works by making a low-rank approximation under the following assumptions:
    - The original term-document matrix is noisy and anecdotal instances of terms are to be eliminated.
      - the approximated matrix is de-noised
    - The original term-document matrix is overly sparse relative to the "true" term-document matrix
      - We want to capture synonymy by reducing the angle between synonymous words
  - Low-rank approximation formed using truncated SVD: $\acute{\mathbf{M}}=\mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^{\mathrm{T}}$
    - No need to actually reconstruct the low-rank approximation; can work equivalently with the result of SVD
      - Rows of $\mathbf{U}_r$ represent r-dimensional term vectors
      - Columns of $\mathbf{V}_r^{\mathrm{T}}$ represent r-dimensional document vectors
      - The columns of $\mathbf{U}_r$ represent *concepts* - literally linear combinations of words
        - Ditto the rows of $\mathbf{V}_r^{\mathrm{T}}$ in which the concepts are linear combinations of documents
    - Can make comparisons:

        - See how related documents j and q are in the low-dimensional space by comparing the vectors $\Sigma_r \mathbf{d}_j$ and $\Sigma_r \mathbf{d}_i$ where $\mathbf{d}_i$ corresponds to the i-th column of $\mathbf{V}_r^T$
            - Typically by cosine similarity
        - Ditto with terms i and p by comparing the vectors $\Sigma_r \mathbf{t}_i$ and $\Sigma_r \mathbf{t}_p$ where $\mathbf{t}_i$ corresponds to the i-th row of $\mathbf{U}_r$.
        - Documents and term vector representations can be clustered using traditional clustering algorithms like k-means using similarity measures like cosine.
    - **Latent Semantic Indexing** allows us to perform searches
        - Given a query, view this as a mini document, and compare it to your documents in the low-dimensional space.
            - Given a query vector $\mathbf{q}$ with dimensionality equal to the number of terms, project it into the document space: $\mathbf{q'} = \Sigma_r^{-1} U_r^T \mathbf{q}$
            - Then compare $\Sigma_r \mathbf{q'}$ against the low-dimensional document vectors $\Sigma_r \mathbf{d}_j$
    - Unfortunately LSA has a few limitations:
        - The resulting dimensions might be difficult to interpret, leading to results which can be justified on the mathematical level, but have no interpretable meaning in natural language.
        - Polysemy isn't captured
        - Word order is ignored
        - The probabilistic model of LSA does not match observed data

Mining semantic correspondences across feature domains

- Extend LSA to working with multiple languages
    - Perform training using LSA procedure with a bilingual (or multilingual) term-document matrix
        - Can perform similarity queries as before, but just using one language (setting all remaining terms to zero irrespective of language they belong to)
    - In the same way you construct queries from a single language, you can create representations for new documents and then append these new vectors to the V matrix so they can be searched
        - Known as **"folding-in"** in the original literature
    - The lower dimensional document vectors for unilingual documents should incorporate the **multilingual synonymy** captured from the training data
    - Known as **Cross-Language LSI** (CL-LSI)
- CL-LSI can be extended further to incorporate data from modalities other than text (and can work with a mixture of modalities)
    - Can be applied to anything that can be represented as vectors that record the **composition** of occurrences in a **unit**
    - For example:
        - represent a set of annotated images by vectors of visual features and occurrences of English words
            - Can perform similarity queries as before
            - Can *project*, or *fold-in* unannotated images
                - Allowing them to be searched
                - Or allowing us to make suggestions of suitable annotations
    - Other uses include:
        - Language modelling
        - Command-based speech recognition

- Spam filtering
    - Pronunciation modelling
- TTS unit selection
- …

Further Reading

- Wikipedia has a good overview of LSA: https://en.wikipedia.org/wiki/Latent_semantic_analysis (https://en.wikipedia.org/wiki/Latent_semantic_analysis)
- The following papers introduce the key techniques
    - Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman (1990). "Indexing by Latent Semantic Analysis" (PDF). Journal of the American Society for Information Science 41 (6): 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9. http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf (http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf)
    - Thomas K. Landauer and Michael L. Littman. Fully Automatic Cross-Language Document Retrieval Using Latent Semantic Indexing. In proc. Sixth Annual Conference of the UW Centre for New Oxford English Dictionary and Text Research. 1990. https://www.cs.rutgers.edu/~mlittman/papers/waterloo91.pdf (https://www.cs.rutgers.edu/~mlittman/papers/waterloo91.pdf)
    - J. R. Bellegarda, "Latent semantic mapping", in IEEE Signal Processing Magazine, vol. 22, no. 5, pp. 70-80, Sept. 2005. doi: 10.1109/MSP.2005.1511825. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1511825 (http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1511825)
    - Hare, Jonathan, Samangooei, Sina, Lewis, Paul and Nixon, Mark (2008) Semantic spaces revisited: investigating the performance of auto-annotation and semantic retrieval using semantic spaces. At CIVR '08: The 2008 international conference on Content-based image and video retrieval, Niagara Falls, Ontario, Canada, 07 - 09 Jul 2008. ACM, 359-368. http://eprints.soton.ac.uk/266160/1/p359.pdf (http://eprints.soton.ac.uk/266160/1/p359.pdf)
    - Hare, Jonathon S., Lewis, Paul H., Enser, Peter G. B. and Sandom, Christine J. (2006) A Linear-Algebraic Technique with an Application in Semantic Image Retrieval. Image and Video Retrieval: 5th International Conference, CIVR 2006, Tempe, AZ, USA, July 2006, LNCS 4, 31-40. http://eprints.soton.ac.uk/262870/1/JH_CIVR2006_Factorisation_Springer.pdf (http://eprints.soton.ac.uk/262870/1/JH_CIVR2006_Factorisation_Springer.pdf)