INF4820

Modeling Word Meaning
Vector Space Models

Erik Velldal

University of Oslo

Oct. 27, 2009

Topics for Today

- Modeling meaning by context
  - Inferring lexical semantics from contextual distributions
  - The distributional hypothesis
  - Ways to define context
  - Frequencies vs. association weights
- Representation in vector space models
  - Feature vectors
  - Feature space
  - Measuring semantic similarity in a “semantic space”

The Distributional Hypothesis

AKA The Contextual Theory of Meaning

- *Meaning is use.* (Wittgenstein, 1953)
- *The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities.* (Harris, 1968)
- *You shall know a word by the company it keeps.* (Firth, 1968)

The Distributional Hypothesis

AKA The Contextual Theory of Meaning

- *Meaning is use.* (Wittgenstein, 1953)
- *The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities.* (Harris, 1968)
- *You shall know a word by the company it keeps.* (Firth, 1968)

> He was feeling seriously hung over after drinking too many shots of **retawerif** at the party last night.

Defining "Context"

- The basic idea: Capture the meaning of a word in terms of its context.
- Motivation: Can compare the meaning of words by comparing their contexts. No need for prior knowledge.
- Each word o_i represented by a set of feature functions $\{f_1, \ldots, f_n\}$. Each f_j records some property of the observed contexts of o_i.
- First task: Define context.

Defining "Context"

- The basic idea: Capture the meaning of a word in terms of its context.
- Motivation: Can compare the meaning of words by comparing their contexts. No need for prior knowledge.
- Each word o_i represented by a set of feature functions $\{f_1, \ldots, f_n\}$. Each f_j records some property of the observed contexts of o_i.
- First task: Define context.

Context windows

- Context = neighborhood of $\pm n$ words before and after the focus word.

Defining "Context"

- The basic idea: Capture the meaning of a word in terms of its context.
- Motivation: Can compare the meaning of words by comparing their contexts. No need for prior knowledge.
- Each word o_i represented by a set of feature functions $\{f_1, \ldots, f_n\}$. Each f_j records some property of the observed contexts of o_i.
- First task: Define context.

Context windows

- Context = neighborhood of $\pm n$ words before and after the focus word.
- Rectangular; treating every word occurring within the window as equally important.
- Triangular; weighting the importance of a context word according to its distance from the target.
- Bag-of-Words (BoW); ignoring the linear ordering of the words.

Defining "Context" (cont'd)

Other BoW Approaches

- Context = all words co-occurring within the same *document*.
- Context = all words co-occurring within the same *sentence*.

Defining "Context" (cont'd)

Other BoW Approaches

- Context = all words co-occurring within the same *document*.
- Context = all words co-occurring within the same *sentence*.

Grammatical relations

- Context = the grammatical relations and dependencies that a target holds to other words.
- Intuition: E.g. nouns occurring in the same grammatical relations with the same verbs probably denote similar kinds of things:

  ... *to* {*drink* | *pour* | *spill*} *some* {*milk* | *water* | *wine*} ...
- Requires deeper linguistic analysis than a simple windowing approach, but PoS-tagging + shallow parsing is enough.

Defining "Context" (cont'd)

What is a word (again)?

- Different levels of abstraction and morphological normalization:
- Full-form words vs. stemming vs. lemmatization ...

Defining "Context" (cont'd)

What is a word (again)?

- Different levels of abstraction and morphological normalization:
- Full-form words vs. stemming vs. lemmatization ...

Stop-words

- Filter out closed-class words or function words by using a so-called stop-list.
- The idea is that only *content* words contributes significantly to indicate the meaning of a word.

Different Types of Contexts ⇒ Different Types of Similarity

- Different kinds of context may indicate different relations of semantic similarity.
- 'Relatedness' *vs.* 'sameness'. Or domain vs. content.
- Similarity in domain :
  {*car, road, gas, service, traffic, driver, license*}
- Similarity in content:
  {*car, train, bicycle, truck, vehicle, airplane, buss*}
- While broader definitions of context (windowing, BoW, etc.) tend to give clues for *domain-based relatedness*, more fine-grained grammatical contexts give clues for *content-based similarity*.

Examples from Oslo Corpus

- Throughout the next lectures we'll sometimes be looking at examples of contextual features extracted from the Oslo Corpus.
- Developed by the Text Laboratory at UiO
- 18.5 mill words
- The corpus is annotated by the Oslo-Bergen Tagger.
- A shallow parser then extracts grammatical features for (lemmatized) *nouns* indicating;
  - adjectival modifications
  - prepositional phrases
  - possessive modification
  - noun-noun conjunction
  - noun-noun modification
  - verbal arguments (subj., dir., ind., and prepositional objects)

Grammatical Context Features

Kunden	*bestilte*	*den*	*mest*	*eksklusive*	*vinen*	*på*	*menyen.*
Customer-the	ordered	the	most	exclusive	wine	on	menu-the.

'The customer ordered the most exclusive wine on the menu.'

- Example of grammatical context features:

Target	Feature
kunde (customer)	`SUBJ_OF bestille` (order)
vin (wine)	`OBJ_OF bestille` (order)
vin (wine)	`ADJ_MOD_BY eksklusiv` (exclusive)
vin (wine)	`PP_MOD_BY meny` (menu)
meny (menu)	`PP_MOD_OF vin` (wine)

Feature Vectors

- A feature vector is an n-dimensional vector of numerical features describing some object.
- Let the set of n feature functions describing the lexical contexts of a word o_i be represented as a feature vector $F(o_i) = \vec{f_i} = \langle f_{i1}, \ldots, f_{in} \rangle$.
- E.g. let $o_i =$ *vin*, and $f_j =$ (OBJ_OF `bestille`).
- Then $f_{ij} = f($*vin*$, ($OBJ_OF `bestille`$)) = 4$ would mean that we have observed *vin* (wine) to be the object of the verb *bestille* (order) in our corpus 4 times.

Feature Vectors

- A feature vector is an n-dimensional vector of numerical features describing some object.
- Let the set of n feature functions describing the lexical contexts of a word o_i be represented as a feature vector $F(o_i) = \vec{f_i} = \langle f_{i1}, \ldots, f_{in} \rangle$.
- E.g. let $o_i = \textit{vin}$, and $f_j = (\text{OBJ\_OF}\ \texttt{bestille})$.
- Then $f_{ij} = f(\textit{vin}, (\text{OBJ\_OF}\ \texttt{bestille})) = 4$ would mean that we have observed *vin* (wine) to be the object of the verb *bestille* (order) in our corpus 4 times.
- A wide range of algorithms for pattern matching and machine learning relies on feature vectors as a means of representing objects numerically.
- (Feature vectors can represent arbitrary objects; e.g. pixels of images for OCR or face recognition.)

The Feature Space

- The feature vectors can be interpreted geometrically; as positioned in a feature space (= vector space model).
- A vector space model is defined by a system of d dimensions or coordinates where objects are represented as real valued vectors in the space $\Re^n$.
- The *dimensions* of our space represent contextual *features*.
- The *points* in our space represent *words* (e.g. noun distributions).
- The points are positioned in the space according to their values along the various contextual dimensions.

Semantic Spaces

- When using a vector space model with context vectors, combined with the distributional hypothesis, we sometimes speak of having defined a semantic space.
- Semantic similarity $\Rightarrow$ Distributional similarity $\Rightarrow$ Spatial proximity

Semantic Spaces

- When using a vector space model with context vectors, combined with the distributional hypothesis, we sometimes speak of having defined a semantic space.
- Semantic similarity $\Rightarrow$ Distributional similarity $\Rightarrow$ Spatial proximity

Formally defined as a triple $\langle \mathrm{F}, A, s \rangle$:

- $\mathrm{F} = \{\vec{f_1}, \ldots, \vec{f_n}\}$ is the set of *feature vectors*. f_{ij} gives the co-occurrence count for the ith word and the jth context.
- A is a *measure of association strength* for a word–context pair, in the form of a statistical test of dependence. Maps each element f_{ij} of the feature vectors in F to a real value.
- s is a *similarity function*.
- (We've talked about F; next up is A, then s.)

Word–Context Association

- We want our feature vectors to reflect which contexts are the most salient or relevant for each word.
- Problem: Raw co-occurrence frequencies alone, or even MLE probabilities, are not a good indicators of relevance.

Word–Context Association

- We want our feature vectors to reflect which contexts are the most salient or relevant for each word.
- Problem: Raw co-occurrence frequencies alone, or even MLE probabilities, are not a good indicators of relevance.
- Consider the noun *vin* (wine) as a direct object of the verbs *kjøpe* (buy) and *helle* (pour):
  - $f(\texttt{vin}, (\texttt{obj\_of kjøpe})) = 14$
  - $f(\texttt{vin}, (\texttt{obj\_of helle})) = 8$
  - ...but the feature `(obj_of helle)` seems more indicative of the semantics of *vin* than `(obj_of kjøpe)`.

Word–Context Association

- We want our feature vectors to reflect which contexts are the most salient or relevant for each word.
- Problem: Raw co-occurrence frequencies alone, or even MLE probabilities, are not a good indicators of relevance.
- Consider the noun *vin* (wine) as a direct object of the verbs *kjøpe* (buy) and *helle* (pour):
  - $f(\texttt{vin}, (\texttt{obj\_of kjøpe})) = 14$
  - $f(\texttt{vin}, (\texttt{obj\_of helle})) = \ 8$
  - ...but the feature `(obj_of helle)` seems more indicative of the semantics of *vin* than `(obj_of kjøpe)`.
- Solution: Weight the frequency counts by an *association function*. "Normalize" frequencies for chance co-occurrence.

Pointwise Mutual Information

- Defines the association between a feature f and an observation o as a likelihood ratio of their joint probability and the product of their marginal probabilities:

$$
\begin{aligned}
I(f,o) &= \log_2 \frac{P(f,o)}{P(f)P(o)} = \log_2 \frac{P(f)P(o|f)}{P(f)P(o)} \\
&= \log_2 \frac{P(o|f)}{P(o)}
\end{aligned}
$$

- Perfect independence: $P(f,o) = P(f)P(o)$ and $I(f,o) = 0$.
- Perfect dependence: If f and o always occur together then $P(o|f) = 1$ and $I(f,o) = \log_2 1/P(o)$.

Pointwise Mutual Information

- Defines the association between a feature f and an observation o as a likelihood ratio of their joint probability and the product of their marginal probabilities:

$$I(f,o) = \log_2 \frac{P(f,o)}{P(f)P(o)} = \log_2 \frac{P(f)P(o|f)}{P(f)P(o)}$$
$$= \log_2 \frac{P(o|f)}{P(o)}$$

- Perfect independence: $P(f,o) = P(f)P(o)$ and $I(f,o) = 0$.
- Perfect dependence: If f and o always occur together then $P(o|f) = 1$ and $I(f,o) = \log_2 1/P(o)$.
- A *smaller marginal probability* $P(o)$ leads to a *larger association score* $I(f,o)$. $\rightarrow$ Overestimates the correlation of rare events.

The Log Odds Ratio

- Measures the magnitude of association between an observed object o and a feature f independently of their marginal probabilities:

$$\log \theta(f, o) = \log \frac{P(f, o)/P(f, \neg o)}{P(\neg f, o)/P(\neg f, \neg o)}$$

- $\theta(f, o)$ expresses how much the chance of observing o increases when the feature f is present.
- $\log \theta(f, o) > 0$ means the probability of seeing o increases when f is present. $\log \theta = 0$ indicates distributional independence.

The Log Odds Ratio

- Measures the magnitude of association between an observed object o and a feature f independently of their marginal probabilities:

$$\log \theta(f, o) = \log \frac{P(f, o)/P(f, \neg o)}{P(\neg f, o)/P(\neg f, \neg o)}$$

- $\theta(f, o)$ expresses how much the chance of observing o increases when the feature f is present.
- $\log \theta(f, o) > 0$ means the probability of seeing o increases when f is present. $\log \theta = 0$ indicates distributional independence.
- There's also a host of other association measures in use, and most take the form of a statistical test of dependence; e.g. the t-test, log likelihood, Fisher's exact test, Jaccard...

Negative Correlations

- Negatively correlated pairs (f, o) are usually ignored when measuring word–context associations (e.g. if $\log \theta(f, o) < 0$).
- Unreliable estimates about negative correlations in sparse data.
- Both unobserved or negatively correlated co-occurrence pairs are assumed to have zero association.

Negative Correlations

- Negatively correlated pairs (f, o) are usually ignored when measuring word–context associations (e.g. if $\log \theta(f, o) < 0$).
- Unreliable estimates about negative correlations in sparse data.
- Both unobserved or negatively correlated co-occurrence pairs are assumed to have zero association.
- We will use $\mathrm{X} = \{\vec{x}_1, \ldots, \vec{x}_k\}$ to denote the set of 'association vectors' that results from applying the association weighting.
- That is, $\vec{x}_i = \langle A(f_{i1}), \ldots, A(f_{in}) \rangle$,
  where $A = \log \theta$

The 20 most salient local contexts of the noun ***teori*** (**theory**):

		Context Feature		
Rank	Frequency	Feat. Type	Feat. Word	Association
0	17	subj_of	forklare (explain, account for)	3.88
1	75	adj_mod_by	økonomisk (economical)	3.74
2	12	adj_mod_by	vitenskapelig (scientific)	3.60
3	5	noun_con	erfaring (experience, practice)	3.30
4	8	obj_of	presentere (present, introduce)	3.25
5	13	obj_of	utvikle (develop, evolve, grow)	3.00
6	6	pp_mod_of	utgangspunkt (point of departure)	2.98
7	5	pp_mod_of	kunnskap (knowledge)	2.81
8	6	adj_mod_by	administrativ (administrative)	2.80
9	4	subj_of	stemme (agree, correspond)	2.71
10	5	subj_of	tilsi (indicate, justify)	2.71
11	5	obj_of	støtte (support, back up,)	2.70
12	6	obj_of	styrke (strengthen)	2.65
13	5	subj_of	beskrive (describe)	2.51
14	4	adj_mod_by	tradisjonell (traditional)	2.49
15	3	subj_of	bekrefte (confirm, acknowledge)	2.44
16	3	subj_of	oppfatte (understand, interpret, perceive)	2.24
17	2	pp_mod_of	motsetning (opposition, opposite, contrast)	2.20
18	3	pp_mod_of	forskjell (difference, distinction)	2.17
19	4	obj_of	nevne (mention)	2.17

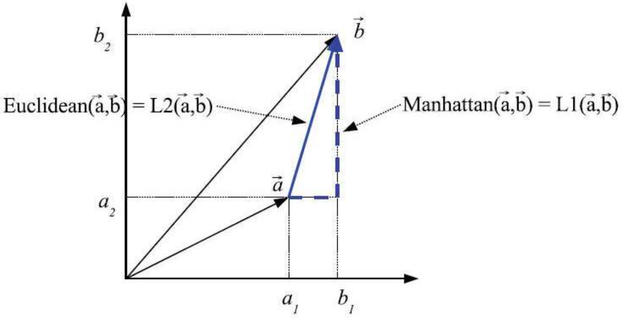
Euclidean Distance

- Vector space models let us compute the *semantic similarity* of words in terms of *spatial proximity*.
- Some standard metrics for measuring *distance* in the space are based on the the family of so-called Minkowski metrics, computing the length (or *norm*) of the *difference* of the vectors;

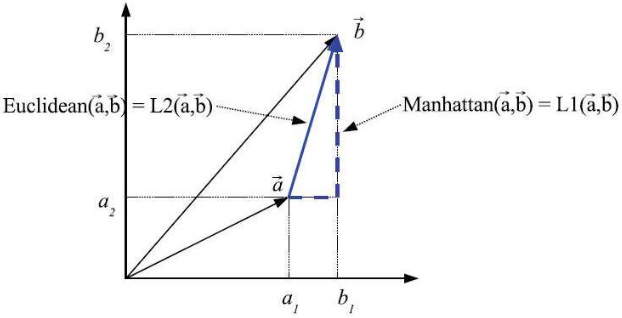
$$d_M(\vec{x}, \vec{y}) = \sqrt[p]{\sum_{i=1}^{n} |\vec{x}_i - \vec{y}_i|^p} \tag{1}$$

- The most commonly used measure is the Euclidean distance or L_2 distance, for which we have $p = 2$
- Other common metrics include the Manhattan distance (or L_1 norm) for which $p = 1$.

Euclidean Distance (cont'd)



Euclidean Distance (cont'd)



- However, a potential problem with the L_2 norm is that it is very sensitive to extreme values and the length of the vectors.
- As vectors of words with different *frequencies* will tend to have different length, the frequency will also affect the similarity judgment.

Euclidean Distance (cont'd)

- Note that, although our association weighting to some degree already 'normalizes' the differences in frequency, words with initially long 'frequency vectors', will also tend to have longer 'association vectors'.

Euclidean Distance (cont'd)

- Note that, although our association weighting to some degree already 'normalizes' the differences in frequency, words with initially long 'frequency vectors', will also tend to have longer 'association vectors'.
- One way to reduce effect of frequency / length is to first normalize all our vectors to have unit length, i.e.:

$$\|\vec{x}\| = \sqrt{\sum_{i=1}^{n} \vec{x}_i^2} = \sum_{i=1}^{n} \vec{x}_i^2 = 1$$

Euclidean Distance (cont'd)

- Note that, although our association weighting to some degree already 'normalizes' the differences in frequency, words with initially long 'frequency vectors', will also tend to have longer 'association vectors'.
- One way to reduce effect of frequency / length is to first normalize all our vectors to have unit length, i.e.:

$$\|\vec{x}\| = \sqrt{\sum\nolimits_{i=1}^{n} \vec{x}_i^2} = \sum_{i=1}^{n} \vec{x}_i^2 = 1$$

- It is also common to instead compute the *cosine* of the angles of the vectors;
  - Under different interpretations the measure is also known as the *normalized correlation coefficient* or the *normalized inner product*...

Cosine Similarity

- Similarity as a function of the angle between the vectors:

$$\cos(\vec{x}, \vec{y}) = \frac{\sum_i \vec{x}_i \vec{y}_i}{\sqrt{\sum_i \vec{x}_i^2} \sqrt{\sum_i \vec{y}_i^2}} = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|}$$

- Constant range between 0 and 1. Avoids the arbitrary scaling caused by dimensionality, frequency or the range of the association measure A.
- As the angle between the vectors shortens, the cosine approaches 1.

Cosine Similarity

- Similarity as a function of the angle between the vectors:

$$\cos(\vec{x}, \vec{y}) = \frac{\sum_i \vec{x}_i \vec{y}_i}{\sqrt{\sum_i \vec{x}_i^2} \sqrt{\sum_i \vec{y}_i^2}} = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|}$$

- Constant range between 0 and 1. Avoids the arbitrary scaling caused by dimensionality, frequency or the range of the association measure A.
- As the angle between the vectors shortens, the cosine approaches 1.
- When applied to *normalized* vectors, the cosine can be simplified to the *dot product* alone:

$$\cos(\vec{x}, \vec{y}) = \vec{x} \cdot \vec{y} = \sum_{i=1}^{n} \vec{x}_i \vec{y}_i$$

Cosine Similarity

- Similarity as a function of the angle between the vectors:

$$\cos(\vec{x}, \vec{y}) = \frac{\sum_i \vec{x}_i \vec{y}_i}{\sqrt{\sum_i \vec{x}_i^2}\sqrt{\sum_i \vec{y}_i^2}} = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\|\|\vec{y}\|}$$

- Constant range between 0 and 1. Avoids the arbitrary scaling caused by dimensionality, frequency or the range of the association measure A.
- As the angle between the vectors shortens, the cosine approaches 1.
- When applied to *normalized* vectors, the cosine can be simplified to the *dot product* alone:

$$\cos(\vec{x}, \vec{y}) = \vec{x} \cdot \vec{y} = \sum_{i=1}^{n} \vec{x}_i \vec{y}_i$$

- The *same relative rank order* as the Euclidean distance for unit vectors.

Next Week

- Computing neighbor relations in the semantic space
- Vector space models for Information Retrieval (IR)
- Representing classes in the vector space
  - Clusters, centroids, memoids...
- Representing class membership
  - Boolean, fuzzy, probabilistic...
- Classification algorithms
  - KNN-classification / c-means, etc.
- Dealing with (very) high-dimensional sparse vectors.
- Reading: The chapter *Vector Space Classification* at `http://informationretrieval.org/`.

Dagan, I., Lee, L., & Pereira, F. (1999). Similarity-based models of word cooccurrence probabilities. *Machine Learning*, *34*(1-3), 43 – 69.

Firth, J. R. (1968). A synopsis of linguistic theory. In F. R. Palmer (Ed.), *Selected papers of j. r. firth: 1952–1959.* Longman.

Grefenstette, G. (1992). SEXTANT: Exploring unexplored contexts for semantic extraction from syntactic analysis. In *Proceedings of the 30th Meeting of the Association for Computational Linguistics* (pp. 324 – 326). Newark, Delaware.

Harris, Z. S. (1968). *Mathematical structures of language.* New York: Wiley.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Acl:90* (pp. 268 – 275). Pittsburgh, USA.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics* (pp. 768 – 774). Montreal, Canada.

Resnik, P. (1993). *Selection and information: A class-based approach to lexical relationships.* Unpublished doctoral dissertation, Department of Computer and Information Science, University of Pennsylvania.

Wittgenstein, L. (1953). *Philosophical investigations.* Oxford: Blackwell.