Contents

Todo list

Chapter 1

Introduction

Computers, machines that play a more and more important role in our lives, require specially designed programming languages to be controlled. This is different from interactions between people where spoken or written natural language is used. Ideally, it would be perfect if interactions with computers was not different to interaction between humans. Computational linguistics is one of the fields that aims to solve this problem.

In order to be controlled by people, or be able to assist people in language related tasks, computers need to understand it. However, different tasks need various level of language "understanding". For instance, even if one does not recognize or know the language of a piece of text on Figure 1.1a, he or she can tell how many words there are, and that there is only one sentence. After a while, one can even say that this is probably a piece of poetry.

The conclusions above require neither deep understanding of the language nor the meaning of the text. Knowledge that texts (at least in some languages) consist of words separated by a space and how poems are usually written is enough. Moreover, knowing the letter combination distribution across the languages or a list of words for all human languages, one would conclude that the text on Figure 1.1a is in Latvian. We managed to get this answers without knowing what the text is about.

On the other hand, a task that provides a list of associations with the text, an essay or a painting inspired by it demands understanding of the text, language knowledge. Luckily, nowadays these kind of problems are not expected or demanded to be done by computers. People enjoy doing these kind of tasks themselves.

However, it is reasonable to ask a computer the following questions regarding text meaning: a) What is the text on Figure 1.1a about? b) What is the relation between the texts on Figure 1.1a and Figure 1.1b? c) Are these texts identical? d) Where did the meeting took place? e) What poems are similar to this?

Text summarizaiton, machine translation, information extraction and retrieval are just a

Jaunkundze ar sunīti

Un Vecrīgas šķērsielā, šaurā
kā vēstuļu kastītes sprauga,
kur troksnim un burzmai tik atbalss,
kur smaržo pēc darvas,
dzelzs un pēc āboliem pagrabos sausos,
es satiku jaunkundzi –
glītu un veiklu kā mēle,
kā spēlējot vijoles lociņš.

(a)

Барышня с собачкой

В Старой Риге, на улице поперечной, узкой,
как щель в почтовый ящик,
в который проникают только отголоски шума, гама,
где запах дёгтя, ржавчины и яблок в сухих подвалах,
я встретил барышню –
красива и ловка - она - язык,
смычок, играющий на скрипке.

(b)

Young Woman with a Dog

On a narrow side-street in Riga's old quarter,
as though in a mailbox slot
where noise and hustle only echo,
and it smells of tar and steel
and apples kept in dry basements,

I met a young woman
attractive and active
as a tongue,
as a violin-bow playing.

(c)

Figure 1.1: Three pieces of written natural language. The text on Figure 1.1a is the beginning of the poem "Jaunkundze ar sunīti" by Aleksandrs Čaks, Figure 1.1b is a translation to Russian by Lora Trin, and Figure 1.1c is an English translation by Inara Cedrins.

few of many branches of computational linguistics that provide methods for answering these questions. The questions above have a general property: all of them are about the meaning of the text. Natural language semantics is an area that studies meaning representation.

Creativity of natural languages—the fact that humans are able to produce and understand sentences they have never came across—complicates meaning modeling. Even if we had a way to map each word to its meaning, it is impractical to apply the same procedure to sentences, because as we process a piece of text most of the sentences in it will be seen for the fist time. Because of this, we need to be able to build the meaning representation of a sentence from its constituent parts: words.

Syntax is a study about the structure of a sentence. Grammars define the rules that are describe how a sentences that belong to a language should look like. For example, a subject is in front of a verb and an object is after it in an English sentence. Having the constituent meaning representation, the meaning of a sentence is built guided by its syntax.

To a first, high level approximation, to be able to deal with the meaning of a text in natural language one needs to have meaning representation of constituents, a view to the (syntactic) structure of the text and a compositional procedure that outputs the meaning representation of the whole text.

Finish introduction

Chapter 2

Overview

2.1 Similarity

Similarity is the degree of resemblance between two objects or events (Hahn, 2014) and plays a crucial role in psychological theories of knowledge and behaviour, where it is used to explain such phenomena as classification and conceptualisation (Tversky, 1977; Tversky and Hutchinson, 1986; Medin et al., 1993; Markman and Gentner, 1996; Hahn and Chater, 1997). *Fruit* is a *category* because it is a practical generalisation. Fruits are sweet and constitute deserts, so when one is presented with an unseen fruit, one can hypothesise that it is served toward the end of a dinner.

Generalisations are extremely powerful in describing a language as well. The verb *runs* requires its subject to be singular. *Verb*, *subject* and *singular* are categories that are used to describe English grammar. When one encounters an unknown word and is told that it is a verb, one will immediately have an idea about how to use it assuming that it is used similarly to other English verbs.

From a computational perspective, this motivates and guides development of similarity components that are embedded into natural language processing systems that deal with tasks such as word sense disambiguation (Schütze, 1998), information retrieval (Salton et al., 1975; Milajevs et al., 2015), machine translation (Dagan et al., 1993), dependency parsing (Hermann and Blunsom, 2013; Andreas and Klein, 2014), dialogue act tagging (Kalchbrenner and Blunsom, 2013; Milajevs and Purver, 2014), reasoning over knowledge bases (Socher et al., 2013), and language modelling (Bengio et al., 2006).

Concretely, a parser might benefit from a generalisation about the part of speech tag of a word which did not occur in the training data based on its occurrence pattern in a large corpus of documents from the web. A dialogue act tagging system, contrastly, might require to classify a whole sentence based on its role in a dialogue. To be useful, the similarity component has to be able to measure similarity between *words* and *sentences*.

According to Hahn (2014) "similarity is an essentially psychological notion, based on the way we represent objects, that is, the way they appear to us." An important consequence of this observation is that before measuring similarity of two entities,[1] their representation has to be obtained.

What's the general term for events and objects? Entity?

2.2 Entity representation for similarity measurement

Section 2.1 established that before measuring similarity, the representation of entities has to be agreed. One needs to be extremely careful when the representation of lexical items is decided, as it is unavoidably connected to the *meaning words in isolation*.

Frege discusses two conflicting principles of meaning (Janssen, 2001). Isolated words meanings are the building blocks of sentence meanings, according to *the principle of compositionality*:

> The meaning of a compound expression is a function of the meaning of its parts and the syntactic rule by which they are combined. (Janssen, 2001)

But the word meaning in isolation is not defined, according to *the principle of contextuality*:

> Never ask for the meaning of a word in isolation, but only in the context of a sentence. (Janssen, 2001)

It worth noting here, that similarity in isolation is also problematic because the number of features an entity has is infinite and its easy to show that two entities will always have an infinite amount of common features (Hahn and Chater, 1997; Goodman, 1972). To make similarity measurement possible, it has to be measured *under a given description* (Hahn, 2014; Medin et al., 1993; Markman and Gentner, 1996), thus similarity is always contextualised. Also, Huth et al. (2016) were able to comprehensively map individual words across cortex, meaning that there are word representations.

Frege's principle of contextuality allows to define the meaning of a word by identifying its contribution to the meaning of a sentence. Firth's (1957) famous quote that "you shall know a word by the company it keeps" suggests that the word meaning can be *modelled* as the combination of the meanings of its occurrences in sentences of a corpus. Note, that this does not provide the absolute word meaning, but only its meaning relative to the corpus. This assumption is also supported by the hypothesis of Harris (1970) that the differences of occurrences of two words *quantify* the difference in their relative meanings.

Once relative word meaning is accepted, compositionality can be used to obtain representations of phrases and sentences (Montague, 1970; Dowty et al., 1980; Janssen, 2016; Coecke et al., 2010; Baroni et al., 2014).

[1] Here and later *entity* is a common term for *events* and *objects*.

2.3 Distributional hypothesis

We would like to capture the intuition that while *John* and *Mary* are distinct, they are rather similar to each other (both of them are humans) and dissimilar to *dog*, *pavement* or *idea*.

Distributional methods provide a way to approach this problem. By representing words and phrases as vectors in a vector space, we can express similarity in meaning via a suitable distance metric within that space.

2.3.1 Co-occurrence based word representation

One way to produce such representations is to directly exploit Harris' (1970) intuition by counting the contexts a word appears in.

For example, one can construct a vector space in which the dimensions correspond to contexts, that are usually other words. The word vector components can then be calculated by taking the frequency with which the word co-occurred with the corresponding contexts within a predefined window in a corpus of interest.

Table 2.1 shows 5 3-dimensional vectors for the words *Mary*, *John*, *girl*, *boy* and *idea*. The words *philosophy*, *book* and *school* label vector space dimensions.

	philosophy	book	school
John	4	60	59
Mary	0	10	22
girl	0	19	93
boy	0	12	146
idea	10	47	39

Table 2.1: Word co-occurrence frequencies extracted from the BNC.

As the vector for *Mary* is closer to *girl* than it is to *boy* in the vector space, we can say that *Mary*'s contexts are similar to *girls*'s (and less similar to *boy*'s), therefore *Mary* is semantically more similar to *girl* than to *boy*.

Is it accurate to talk about vector closeness?

Mathematically the similarity can be expressed using, for instance, the cosine of the angle between two vectors:

$$\cos(\theta) = \frac{\overrightarrow{Mary} \cdot \overrightarrow{girl}}{||\overrightarrow{Mary}||\overrightarrow{girl}||} = \frac{0 \times 0 + 10 \times 19 + 22 \times 93}{\sqrt{0^2 + 10^2 + 22^2}\sqrt{0^2 + 19^2 + 93^2}} \approx \frac{2236}{2294} \approx 0.975$$

$$\cos(\phi) = \frac{\overrightarrow{Mary} \cdot \overrightarrow{boy}}{||\overrightarrow{Mary}||\overrightarrow{boy}||} = \frac{0 \times 0 + 10 \times 12 + 22 \times 146}{\sqrt{0^2 + 10^2 + 22^2}\sqrt{0^2 + 12^2 + 146^2}} \approx \frac{3332}{3540} \approx 0.941$$

where θ is the angle between the vectors of *Mary* and *girl*; and ϕ is the angle between the vectors of *Mary* and *boy*.

In the current example of a naïve example vector space, *John* is also closer to *girl* than to *boy*, which is counter-intuitive. This might be because of the small number of dimen-

sions used, the poor selection of the context words, or the usage of raw co-occurrence numbers. Refer to Turney and Pantel (2010) and Levy et al. (2015) for the discussion of vector space parameters, and see Kiela and Clark (2014), Lapesa and Evert (2014) and below for a detailed comparison of their tuning and performance.

reference the section about parameters

2.3.2 Neural word embedding

Deep learning techniques use this distributional hypothesis differently. Instead of relying on observed co-occurrence frequencies, a neural model is trained to maximise some objective function related to e.g. the probability of observing the surrounding words in some context (Mikolov et al., 2013a):

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \tag{2.1}$$

Maximising this function produces vectors which maximise the conditional probability of observing words in a context around the target word w_t, where c is the size of the training context, and $w_1 w_2, \cdots w_T$ is a sequence of training words. They therefore capture the distributional intuition and can express degrees of lexical similarity.

However, they have also proved successful at other tasks (Mikolov et al., 2013b); the vectors obtained encode not only attributional similarity (similar words are close to each other), but also relational similarities (**?**). For example, it is possible to extract the singular:plural relation (*apple*:*apples*, *car*:*cars*) using vector subtraction:

$$\overrightarrow{apple} - \overrightarrow{apples} \approx \overrightarrow{car} - \overrightarrow{cars}$$

also semantic relationships are preserved:

$$\overrightarrow{king} - \overrightarrow{man} \approx \overrightarrow{queen} - \overrightarrow{woman}$$

allowing the formation of analogy queries similar to $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} = ?$, obtaining $\overrightarrow{queen}$ as the result.[2]

Both neural and co-occurrence-based approaches have advantages over classical formal approaches in their ability to capture lexical semantics and degrees of similarity; their success at extending this to the sentence level, and to more complex semantic phenomena, depends on their applicability within compositional models.

[2] Levy et al. (2014) improved Mikolov et al. (2013b)'s method of retrieving relational similarities by changing the objective function and improved the state-of-the-art results both for neural embeddings and co-occurrence based vectors.

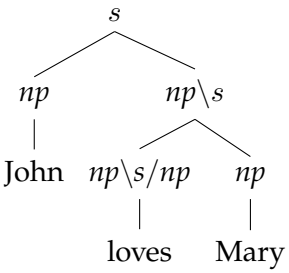

Figure 2.1: A syntactic tree for *John loves Mary*. Lexicon assigns categories to words: *John* is *np*, loves is *np**s*/*np* and Mary is *np*. Backward and forward composition rules derive the syntactic tree.

2.4 Formal semantics

Formal approaches to the semantics of natural language have long built upon the classical idea of compositionality—that the meaning of a sentence is a function of its parts (Janssen, 2001). In compositional type-logical approaches, predicate-argument structures representing phrases and sentences are built from their constituent parts by general operations such as beta-reduction within the lambda calculus (Montague, 1970): for example, given a semantic representation of *John* as $john'$ and *loves* as $\lambda y.\lambda x.loves'(x, y)$, the sentence *John loves Mary* can be constructed as

$$\lambda y.\lambda x.loves'(x, y)(mary')(john') = loves'(john', mary')$$

To get the semantic representation of the sentence *John loves Mary* we need to do the following. Syntactic rules define how constituents are combined to form other constituents (and finally a sentence). Translation rules define how semantic representations of the constituents are combined to get a semantic representation of the whole.

Categorial grammars are widely used to obtain syntactic structure of a sentence. Given a set of basic categories `ATOM`, for example $\{n, s, np\}$ complex categories `CAT`\`CAT` and `CAT`/`CAT` can be constructed, where `CAT` is either an element of `ATOM` or a complex category. So the transitive verb category is `np`\`s`/`np`. Intuitively we want to say that obtaining a sentence with a transitive verb there must be a noun phrase before and after it.

Parsing is done by composing categories together according to two rules:

1. **Backward application**: If α is a string of category A and β is a string of category $A\backslash B$, then $\alpha\beta$ is of category B.

2. **Forward application**: If α is a string of category A and β is a string of category B/A, then $\beta\alpha$ is of category B.

Figure 2.1 illustrates the parse tree for *John loves Mary* obtained using the category composition rules.

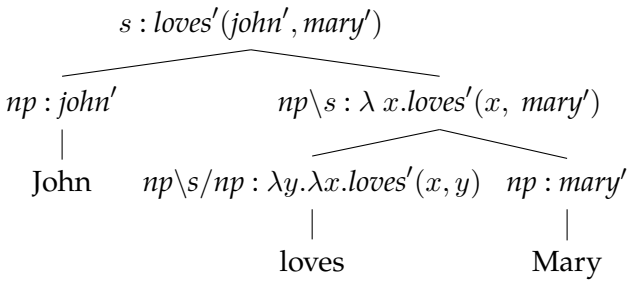

Figure 2.2: A syntactic tree for *John loves Mary*. Lexicon assigns categories to words: *John* is *np*, loves is $np\backslash s/np$ and Mary is *np*. Backward and forward composition rules derive the syntactic tree.

The last step is to map syntactic categories with semantic terms. Again, there are base types (e for entities and t for sentences) and complex types of the form $(a \rightarrow b)$ where a and b are types. The mapping between syntactic categories and semantic types is defined as a function *type*:

$$
\begin{aligned}
&type(np) = e \\
&type(s) = t \\
&type(A/B) = (type(B) \rightarrow type(A)) \\
&type(B\backslash A) = (type(B) \rightarrow type(A))
\end{aligned}
$$

Syntactic backward and forward application corresponds to functional application. The final result would be the this:

Given a suitable pairing between a syntactic grammar, semantic representations and corresponding general combinatory operators, this can produce structured sentential representations with broad coverage and good generalisability (Bos, 2008). This logical approach is extremely powerful because it can capture complex aspects of meaning such as quantifiers and their interaction (Copestake et al., 2005)), and enables inference using well studied and developed logical methods (Bos and Gabsdil, 2000).

2.5 Composition

Methods based on this distributional hypothesis have recently been applied to many tasks, but mostly at the word level: for instance, word sense disambiguation (**?**) and lexical substitution (**?**). They exploit the notion of similarity which correlates with the angle between word vectors (**?**). *Compositional* distributional semantics goes beyond the word level and models the meaning of phrases or sentences based on their parts. **?** perform composition of word vectors using vector addition and multiplication operations. The limitation of this approach is the operator associativity, which ignores the argu-

ment order, and thus word order. As a result, "*John loves Mary*" and "*Mary loves John*" get assigned the same meaning.

Concretely, if *John*, *Mary* and *loves* meaning is represented as vectors $\overrightarrow{john}$, $\overrightarrow{mary}$ and $\overrightarrow{loves}$, the meaning of the sentence *John loves Mary* is $\overrightarrow{john} + \overrightarrow{loves} + \overrightarrow{mary}$.

The To capture word order, various approaches have been proposed. **?** extend the compositional approach by using non-associative linear algebra operators as proposed in the theoretical work of (**?**).

The functional application of semantic term can be replaced with tensors (**?**). Then, a transitive verb is represented by matrix, which can be obtained from a corpus using the formula $\sum_i \overrightarrow{s_i} \otimes \overrightarrow{o_i}$ (the relation method of (**?**)), where $\overrightarrow{s_i}$ and $\overrightarrow{i_i}$ are the subject–object pairs of the verb.

The vector of the whole sentence is $\overline{loves} \odot (\overrightarrow{john} \otimes \overrightarrow{mary})$.

2.6 The tasks that require similarity

Dialogue act tagging There are many ways to approach the task of dialogue act tagging (**?**). The most successful approaches combine *intra*-utterance features, such as the (sequences of) words and intonational contours used, together with *inter*-utterance features, such as the sequence of utterance tags being used previously. To capture both of these aspects, sequence models such as Hidden Markov Models are widely used (**??**). The sequence of words is an observable variable, while the sequence of dialogue act tags is a hidden variable.

However, some approaches have shown competitive results without exploiting features of inter-utterance context. **?** concentrate only on features found inside an utterance, identifying ngrams that correlate strongly with particular utterance tags, and propose a statistical model for prediction which produces close to the state of the art results.

The current state of the art (**?**) uses Recurrent Convolutional Neural Networks to achieve high accuracy. This model includes information about word identity, intra-utterance word sequence, and inter-utterance tag sequence, by using a vector space model of words with a compositional approach. The words vectors are not based on distributional frequencies in this case, however, but on randomly initialised vectors, with the model trained on a specific corpus. This raises several questions: what is the contribution of word sequence and/or utterance (tag) sequence; and might further gains be made by exploiting the distributional hypothesis? What is the contribution of utterance meaning to its tag?

Paraphrase detection Microsoft paraphrase corpus (**?**) is a collection of sentences labeled whether one is a paraphrase of another.

Disambiguation The transitive verb disambiguation dataset[3] described in (**?**). The dataset consists of ambiguous transitive verbs together with their arguments; landmark verbs, which identify one of the verb senses; and human judgements which specify the similarity to the landmarks of the disambiguated sense of the verb in the context given. This is similar to the intransitive dataset described in (**?**).

Consider the sentence *"system meets specification"*; here, *meets* is the ambiguous transitive verb, and *system* and *specification* are the arguments in this context. Possible landmarks for *meet* are *satisfy* and *visit*; for this sentence, the human judgements show that the disambiguated verb meaning is similar to the landmark *satisfy*, and less similar to *visit*.

The task is to estimate the similarity of the sense of a verb in a context with a given landmark. To provide our similarity measures, we compose the verb with its arguments using one of our compositional operators, do the same for the landmark and the arguments, and compute the cosine similarity of the two vectors. To evaluate performance, we average the human judgements for the same verb, argument and landmark entries, and use the average values to calculate the correlation. As a baseline, we compare to the correlation achieved using only the verb vector, without composing with its arguments.

Sentence similarity The transitive sentence similarity dataset described in (**?**). The dataset consists of transitive sentence pairs and a human similarity judgement. The task is to estimate similarity between two sentences.

[3]This and the sentence similarity datasets are available at http://www.cs.ox.ac.uk/activities/compdistmeaning/

Bibliography

Ulrike Hahn. Similarity. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(3):271–280, 2014. ISSN 1939-5086. doi: 10.1002/wcs.1282. URL http://dx.doi.org/10.1002/wcs.1282.

Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977. doi: 10.1037/0033-295x.84.4.327. URL http://dx.doi.org/10.1037/0033-295X.84.4.327.

Amos Tversky and J. Wesley Hutchinson. Nearest neighbor analysis of psychological spaces. *Psychological Review*, 93(1):3 – 22, 1986. ISSN 0033-295X. URL http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=1986-13502-001&site=ehost-live.

Douglas L Medin, Robert L Goldstone, and Dedre Gentner. Respects for similarity. *Psychological review*, 100(2):254, 1993. doi: 10.1037/0033-295X.100.2.254. URL http://psycnet.apa.org/journals/rev/100/2/254/.

Arthur B. Markman and Dedre Gentner. Commonalities and differences in similarity comparisons. *Memory & Cognition*, 24(2):235–249, 1996. ISSN 1532-5946. doi: 10.3758/BF03200884. URL http://dx.doi.org/10.3758/BF03200884.

Ulrike Hahn and Nick Chater. Concepts and similarity. *Knowledge, concepts and categories*, pages 43–92, 1997. URL https://books.google.co.uk/books?hl=en&lr=&id=pc3XAQAAQBAJ&oi=fnd&pg=PA43&dq=Concepts+and+similarity.&ots=it4aC91-qv&sig=ebsOGf72FEIpE0WX6HZoWhi2SsQ#v=onepage&q=Concepts%20and%20similarity.&f=false.

Hinrich Schütze. Automatic word sense discrimination. *Comput. Linguist.*, 24(1):97–123, March 1998. ISSN 0891-2017. URL http://dl.acm.org/citation.cfm?id=972719.972724.

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975. ISSN 0001-0782. doi: 10.1145/361219.361220. URL http://doi.acm.org/10.1145/361219.361220.

Dmitrijs Milajevs, Mehrnoosh Sadrzadeh, and Thomas Roelleke. Ir meets nlp: On the semantic similarity between subject-verb-object phrases. In *Proceedings of the 2015 International Conference on Theory of Information Retrieval*, ICTIR '15, pages 231–240, New

York, NY, USA, 2015. ACM. ISBN 978-1-4503-3833-2. doi: 10.1145/2808194.2809448. URL http://www.eecs.qmul.ac.uk/~dm303/static/ictir006-milajevs.pdf.

Ido Dagan, Shaul Marcus, and Shaul Markovitch. Contextual word similarity and estimation from sparse data. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, pages 164–171, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. doi: 10.3115/981574.981596. URL http://dx.doi.org/10.3115/981574.981596.

Karl Moritz Hermann and Phil Blunsom. The role of syntax in vector space models of compositional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 894–904, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P13-1088.

Jacob Andreas and Dan Klein. How much do word embeddings encode about syntax? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 822–827, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-2133.

Nal Kalchbrenner and Phil Blunsom. Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 119–126, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-3214.

Dmitrijs Milajevs and Matthew Purver. Investigating the contribution of distributional semantic information for dialogue act classification. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 40–47, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-1505.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 926–934. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5028-reasoning-with-neural-tensor-networks-for-knowledge-base-completion.pdf.

Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006. URL http://machinelearning.wustl.edu/mlpapers/paper_files/BengioDVJ03.pdf.

Theo M.V. Janssen. Frege, contextuality and compositionality. *Journal of Logic, Language and Information*, 10(1):115–136, 2001. ISSN 1572-9583. doi: 10.1023/A:1026542332224. URL http://dx.doi.org/10.1023/A:1026542332224.

Nelson Goodman. Problems and projects. 1972.

Alexander G Huth, Wendy a De Heer, Thomas L Griffiths, Frédéric E Theunissen, and L Jack. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016. ISSN 0028-0836. doi: 10.1038/nature17637. URL http://dx.doi.org/10.1038/nature17637.

John R. Firth. A Synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis*, pages 1–32, 1957.

Zellig S. Harris. *Papers in Structural and Transformational Linguistics*, chapter Distributional Structure, pages 775–794. Springer Netherlands, Dordrecht, 1970. ISBN 978-94-017-6059-1. doi: 10.1007/978-94-017-6059-1_36. URL http://dx.doi.org/10.1007/978-94-017-6059-1_36.

Richard Montague. Universal grammar. *Theoria*, 36(3):373–398, 1970. ISSN 1755-2567. doi: 10.1111/j.1755-2567.1970.tb00434.x. URL http://dx.doi.org/10.1111/j.1755-2567.1970.tb00434.x.

David R. Dowty, Robert E. Wall, and Stanley Peters. *Introduction to Montague Semantics*. Springer Netherlands, 1980. doi: 10.1007/978-94-009-9065-4. URL http://dx.doi.org/10.1007/978-94-009-9065-4.

Theo M. V. Janssen. Montague semantics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2016 edition, 2016. URL http://plato.stanford.edu/archives/spr2016/entries/montague-semantics/.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *CoRR*, abs/1003.4394, 2010. URL http://arxiv.org/abs/1003.4394.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. Frege in space: A program of compositional distributional semantics. *LiLT (Linguistic Issues in Language Technology)*, 9, 2014. URL http://csli-lilt.stanford.edu/ojs/index.php/LiLT/article/view/6.

Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, January 2010. ISSN 1076-9757. URL http://arxiv.org/pdf/1003.1141v1.pdf.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015. ISSN 2307-387X. URL https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/570.

Douwe Kiela and Stephen Clark. A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their*

Compositionality (CVSC), pages 21–30, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-1503.

Gabriella Lapesa and Stefan Evert. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2:531–545, 2014. URL http://www.aclweb.org/anthology/Q14-1041.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013a. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, pages 746–751, 2013b. URL http://www.aclweb.org/anthology/N13-1090.pdf.

Omer Levy, Yoav Goldberg, and Israel Ramat-Gan. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Baltimore, Maryland, USA, June. Association for Computational Linguistics*, 2014. URL http://www.cs.bgu.ac.il/~yoavg/publications/conll2014analogies.pdf.

Johan Bos. Wide-Coverage Semantic Analysis with Boxer. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 277–286. College Publications, 2008. URL http://www.aclweb.org/anthology/W08-2222.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3(2):281–332, 2005. ISSN 1572-8706. doi: 10.1007/s11168-006-6327-9. URL http://dx.doi.org/10.1007/s11168-006-6327-9.

Johan Bos and Malte Gabsdil. First-order inference and the interpretation of questions and answers. *Proceedings of Gotelog*, pages 43–50, 2000.