Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018)

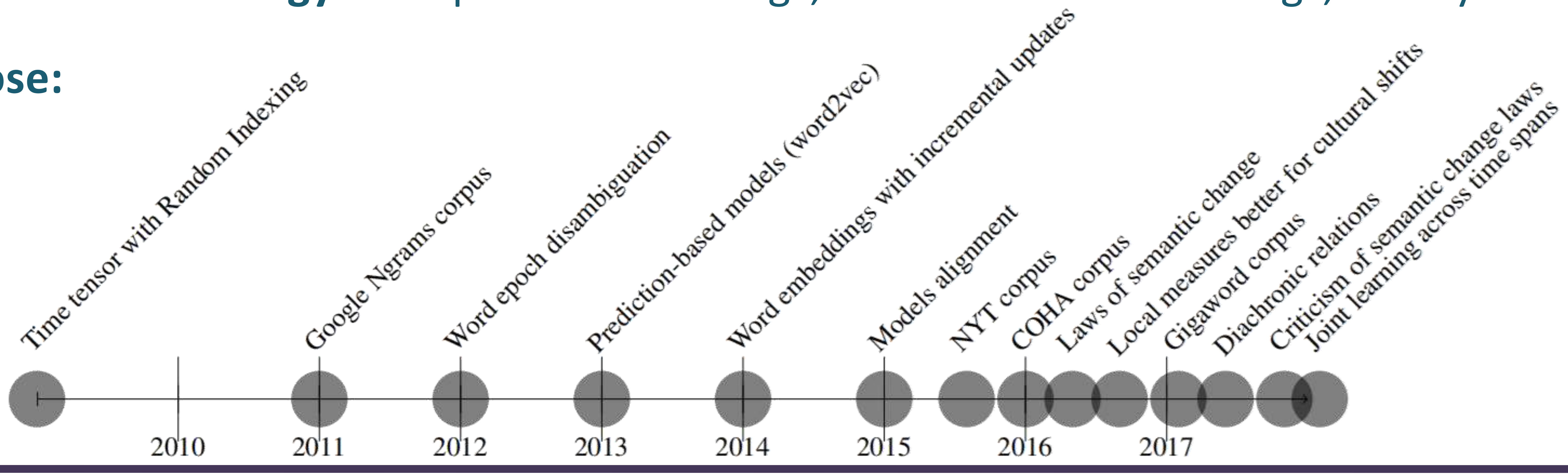
Diachronic word embeddings and semantic shifts: a survey

OUTLINE

- Introduction
- The concept of semantic shifts
- Tracing semantic shifts distributionally
- Laws of semantic change
- Diachronic semantic relations
- Applications
- Open challenges

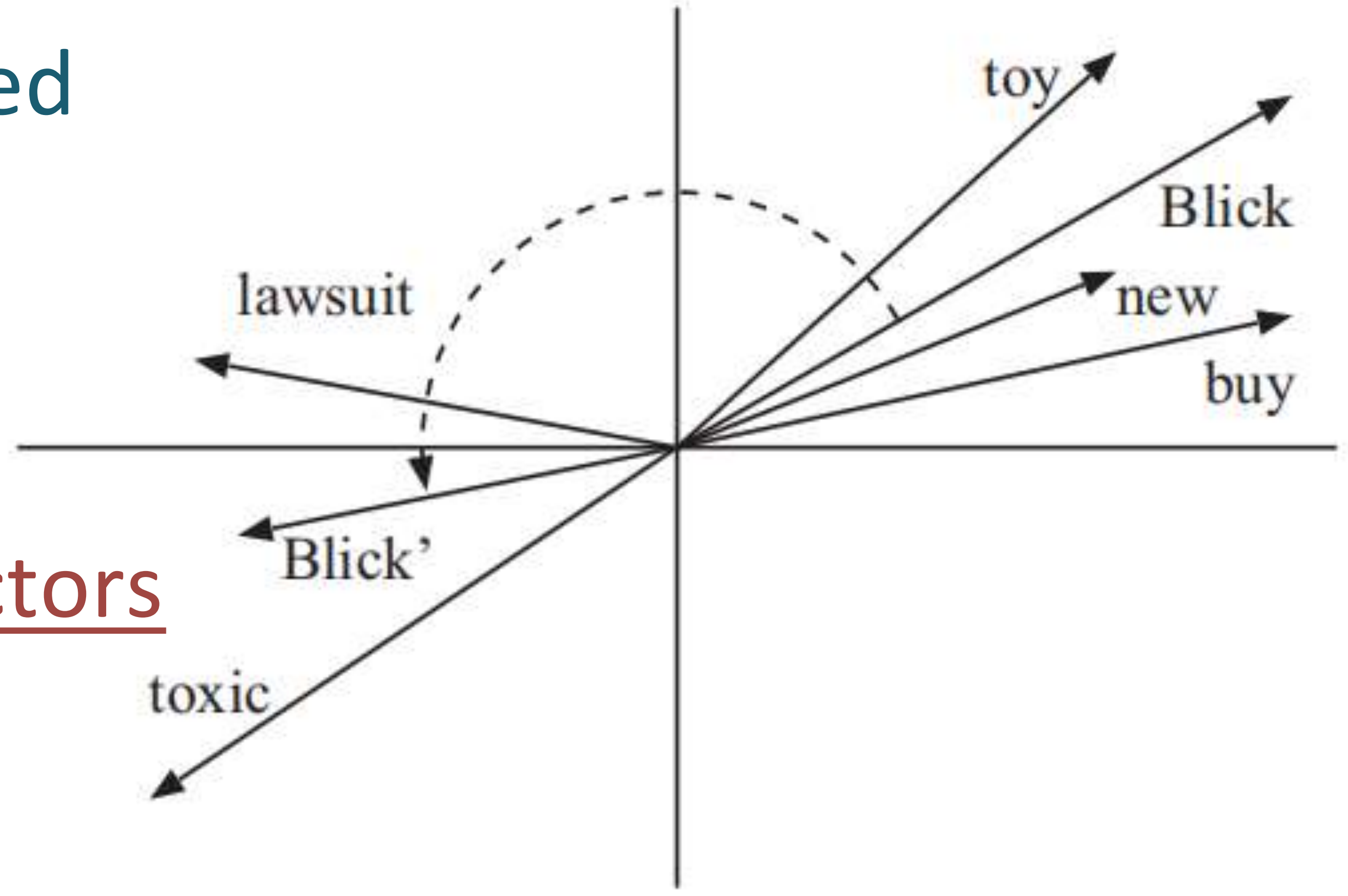
Introduction

- **Diachronic Semantic Shifts: the meanings of words continuously change over time**
  - Changes to the core meaning of words (*gay*: ‘carefree’ → ‘homosexual’ during the 20th century)
  - Subtle shifts of cultural associations (*Iraq* or *Syria*: → associated with the concept of ‘war’ after conflicts)
- **Word Embeddings: a vector that can approximate meaning and represent a word in a lower dimensional space**
  - **Interested research communities**: natural language processing, information retrieval, and political science
  - **Unstandardized terminology**: ‘temporal embeddings,’ ‘diachronic embeddings,’ or ‘dynamic embeddings’
- **Research Purpose:**



Introduction

- **Scope of Research:**
  - **Restricted to**: semantic shifts using distributional word embedding models (that is, representing lexical meaning with dense vectors produced from co-occurrence data)
    - **Distributed hypothesis**: For two words, their similarity in meaning is predicted by the similarity of their distributions of co-occurring words
    - As Firth puts it, “You shall know a word by the company it keeps.”
    - In a semantic space, a word’s semantics are mapped to high dimensional vectors in a geometric space. The dimensions of the space represent distinctions between the meanings of words
      - Accordingly, words with similar semantics have similar vector representations
  - **Only briefly mention**: other data-driven approaches also employed to analyze temporal-labeled corpora (for example, topic modeling)
  - **Not cover**: syntactic shifts and other changes in the functions rather than meaning



The Concepts of Semantic Shifts

- **Theoretical (diachronic) linguistics: focuses on the study of how to express regularities of linguistic change**
  - **Phonology**: accounts for changes in the linguistic sound system → laws of phonological change (e.g. Grimm's law, the great vowel shift)
  - **Lexical semantics**: studies the evolution of word meaning over time
    - **Semantic shifts**: innovations changing the lexical meaning rather than the grammatical function of a form
- **Much theoretical work has been devoted to documenting and categorizing various types of semantic shifts**
  - The categorization found in Bloomfield (1933) is arguably the most used and has inspired recent studies
  - originally proposed nine classes of semantic shifts, six of which are complimentary pairs along a dimension
    - 'narrowing' – 'broadening' (e.g. Old English *mete* 'food' → English *meat* 'edible flesh')
    - speaker's attitude as one of either degeneration or elevation (e.g. Old English *cniht* 'boy, servant' → *knight*)

The Concepts of Semantic Shifts

- **The categorization found in Bloomfield (1933)**
  - **Narrowing**: Change from superordinate level to subordinate level, e.g. *skyline* → "decorated by skyscrapers"
  - **Widening**: Generonyms or genericization, e.g. *Kleenex* "specific brand name" → "the product, facial tissue"
  - **Metaphor**: Change based on similarity of thing, e.g. *broadcast* originally meant "to cast seeds out"
  - **Metonymy**: Change based on nearness in space or time, e.g., *jaw* "cheek" → "mandible"
  - **Synecdoche**: Change based on whole-part relation, e.g. capital cities = countries or their governments
  - **Hyperbole**: Change from weaker to stronger meaning, e.g., *kill* "torment" → "slaughter"
  - **Meiosis**: Change from stronger to weaker meaning, e.g., *astound* "strike with thunder" → "surprise strongly"
  - **Degeneration**: E.g., *knave* "boy" → "servant" → "deceitful or despicable man"; *awful* "awe-inspiring" → "bad"
  - **Elevation**: E.g., *knight* "boy" → "nobleman"; *terrific* "terrifying" → "astonishing" → "very good"

The Concepts of Semantic Shifts

- **Why do semantic shifts occur?**
  - **Driving forces**
    - **Linguistic causes** (by the process of ellipsis, or by the need for discrimination of synonyms)
    - **Psychological causes** (by changes in the attitudes of speakers)
    - **Sociocultural causes** (by changes in the general environment of the speakers)
    - **Cultural/encyclopedic causes**
  - **Categorization**
    - **Linguistic drifts**: Slow and regular changes in core meaning of words
    - **Cultural shifts**: culturally determined changes in associations of a given word
    - ***Substitution***: by changes of technological process. e.g. *car* “non-motorized vehicles” → “automobile”

The Concepts of Semantic Shifts

- **Availability of large corpora enabled the development of new methodologies for lexical semantic shifts**
  - **Key assumption: changes in a word's collocational patterns reflect changes in word meaning (Hilpert, 2008)**
    - Kerremans et al. (2010) study *detweet*, showing the two separate usages/meanings for this word ('to delete from twitter,' vs 'to avoid tweeting') based on large amounts of web-crawled data.
  - The usage-based view of lexical semantics aligns well with the assumptions underlying the distributional semantic approach (Firth, 1957) often employed in NLP
    - The time spans studied are often considerably shorter (decades, rather than centuries) and these distributional methods are well suited for monitoring the gradual process of meaning change
    - Gulordava and Baroni (2011), for instance, showed that distributional models capture cultural shifts, like the word *sleep* acquiring more negative connotations related to sleep disorders, when comparing its 1960s contexts to its 1990s contexts
  - Semantic shifts are often reflected in large corpora through change in the context of the word which is undergoing a shift, as measured by co-occurring words

Tracing Semantic Shifts Distributionally

- **The task of discovery of semantic shifts from data**
  - Given corpora [C_1, C_2, ...C_n] containing texts created in time periods [1, 2, ...n], the task is to locate words with different meaning in different periods, or to locate the words which changed most. Other tasks are possible: discovering general trends in semantic shifts or tracing the dynamics of the relationships between words
- **Sources of diachronic data for training and testing**
  - When automatically detecting semantic shifts, the types of generalizations we will be able to infer are influenced by properties of the textual data being used, such as the source of the datasets and the temporal granularity of the data
- **Training data**
  - Time unit (granularity of temporal dimension) can be chosen before slicing the text collection into subcorpora
    - Earlier works dealt with long-term ones (spanning decades or even centuries), as they are easier to trace
    - Sagi et al. (2011) study differences between Early Middle, Late Middle, and Early Modern English

Tracing Semantic Shifts Distributionally

- **The release of the Google Books Ngrams corpus**
  - **Significance**: Promoted the development of long-term semantic shifts and spurred work on ‘culturomics’
  - **Examples**: Mihalcea and Nastase (2012) used this dataset to detect differences in word usage and meaning across 50-years time spans, while Gulordava and Baroni (2011) compared word meanings in the 1960s and in the 1990s, achieving good correlation with human judgments
  - **Limitations**: Not contain full texts. However, for many cases, this corpus was enough, and its usage as the source of diachronic data continued in Mitra etal. (2014) (employing syntactic ngrams), who detected word sense changes over several different time periods spanning from 3 to 200 years
- **New research prior to the publication of this paper**
  - **Difference**: In general, corpora with smaller time spans are useful for analyzing socio-cultural semantic shifts, while corpora with longer spans are necessary for the study of linguistically motivated semantic shifts
  - **Characterization**: Time spans tend to decrease in size and become more granular, as researchers are attempting to trace increasingly subtle cultural semantic shifts (more relevant for practical tasks)

Tracing Semantic Shifts Distributionally

- **Examples of new research prior to the publication of this paper**

  - Kim et al. (2014) and Liao and Cheng (2016) analyzed the *yearly* changes of words. Note that, instead of using granular 'bins', time can also be represented as a continuous differentiable value (Rosenfeld and Erk, 2018)

  - Kulkarni et al. (2015) used Amazon Movie Reviews (with granularity of 1 year) and Twitter data (with granularity of 1 month). Their results indicated that computational methods for the detection of semantic shifts can be robustly applied to time spans less than a decade

  - Another *yearly* text collection, the New-York Times Annotated Corpus (Sandhaus, 2008)

    - Zhang et al. (2015) again managed to trace subtle semantic shifts. The same corpus was employed by Szymanski (2017), with 21 separate models, one for each year from 1987 to 2007, and to some extent by Yao et al. (2018), who crawled the NYT web site to get 27 *yearly* subcorpora (from 1990 to 2016).

  - Corpus of Historical American (COHA)

    - The inventory of diachronic corpora used in tracing semantic shifts was expanded by Eger and Mehler (2016) with time slices equal to one decade

Tracing Semantic Shifts Distributionally

- **Test Sets/Evaluation Strategies**
  - Diachronic corpora are needed not only as a source of *training* data for developing semantic shift detection systems, but also as a source of *test* sets to evaluate such systems - more complicated
  - **Ideal Situation**: Diachronic approaches should be evaluated on human-annotated lists of semantically shifted words (ranked by the degree of the shift)
  - **Difficulty**: such gold standard data is difficult to obtain, even for English, let alone for other languages
  - **Solutions**
    - **Word epoch disambiguation**: Mihalcea and Nastase (2012) evaluated the ability of a system to detect the time span that specific contexts of a word undergoing a shift belong to
    - **Cross-time alignment**: a system has to find equivalents for certain words in different time periods (for example, 'Obama' in 2015 corresponds to 'Trump' in 2017)
    - use the detected diachronic semantic shifts to trace or predict real-world events like armed conflicts

Tracing Semantic Shifts Distributionally

- **Requirement:** all these evaluation methods still require the existence of large manually annotated semantic shift datasets. The work to properly create and curate such datasets is in its infancy
  - **One approach to avoid this requirement:** The approach is borrowed from research on word sense disambiguation and consists of making a synthetic task by merging two real words together and then modifying the training and test data according to a predefined sense-shifting function
    - Rosenfeld and Erk (2018) successfully employed this approach to evaluate their system
      - The real words are randomly sampled from two distinct semantic classes from the BLESS dataset
      - Banana + lobster = banana ◦ lobster, researchers can capture how similar banana ◦ lobster is to banana by comparing banana ◦ lobster to words in the fruit BLESS class
    - **The disadvantage of this method**: it still operates on synthetic words, limiting the ability of this evaluation scheme to measure the models' performance with regards to real semantic shift data
  - **Thus**: the problem of evaluating semantic shift detection approaches is far from being solved, and practitioners often rely on self-created test sets, or even simply manually inspecting the results

Tracing Semantic Shifts Distributionally

- **Methodology of extracting semantic shifts from data**

  - Before the broad adoption of word embedding models, it was quite common to use change in raw word frequencies in order to trace semantic shifts or other kinds of linguistic change

    - Researchers also studied the increase or decrease in the frequency of a word A collocating with another word B over time, and based on this inferred changes in the meaning of A. However, it is clear that semantic shifts ≠ changes in word frequency (or this connection may be very subtle and non-direct)

  - Such an approach should be superior to frequency-proxied methods. A number of recent publications have showed that *distributional word representations* provide an efficient way to solve these tasks

    - They represent meaning with sparse or dense vectors, produced from word co-occurrence counts. Although conceptually the source of the data for these models is still word frequencies, they 'compress' this information into continuous lexical representations which are both efficient and convenient to work with.

    - Kulkarni et al. (2015) explicitly demonstrated that distributional models outperform the frequency-based methods in detecting semantic shifts. They managed to trace semantic shifts more precisely and with greater explanatory power (*gay*: through time, its nearest semantic neighbors changed)

Tracing Semantic Shifts Distributionally

- **Distributional models were being used indiachronic research long before the paper of Kulkarni et al. (2015)**

  - But there was no rigorous comparison to the frequentist methods. Already in 2009, it was proposed that one can use distributional methods to detect semantic shifts in a quantitative way.

    - The pioneering work by Jurgens and Stevens (2009) described an insightful conceptualization of a sequence of distributional model updates through time: it is effectively a Word:Semantic Vector:Time tensor, in the sense that each word in a distributional model possesses a set of semantic vectors for each time span we are interested in. They employed the *Random Indexing* (RI) algorithm to create word vectors

  - Two years later

    - Gulordava and Baroni (2011) used explicit count-based models, consisting of sparse co-occurrence matrices weighted by Local Mutual Information, while Sagi et al. (2011) turned to Latent Semantic Analysis

    - In Basile et al. (2014), an extension to RI dubbed *Temporal Random Indexing* (TRI) was proposed. However, no quantitative evaluation of this approach was offered (only a few hand-picked examples based on the Italian texts from the *Gutenberg Project*), and thus it is unclear whether TRI is any better than other distributional models for the task of semantic shift detection

Tracing Semantic Shifts Distributionally

- **Further on, the diversity of the employed methods started to increase**

  - Mitra et al. (2014) analyzed clusters of the word similarity graph in the subcorpora corresponding to different time periods. Their distributional model consisted of lexical nodes in the graphs connected with weighted edges. The weights corresponded to the number of shared most salient syntactic dependency contexts

    - Importantly, they were able to detect not only the mere fact of a semantic shift, but also its type. Thus, this work goes into a much less represented class of 'fine-grained' approaches to semantic shift detection. It is also important that Mitra et al. (2014) handle natively the issue of polysemous words, putting the much-neglected problem of word senses in the spotlight.

  - The work of Kim et al. (2014) was seminal in the sense that it is arguably the first one employing prediction-based word embedding models to trace diachronic semantic shifts. Particularly, they used incremental updates and Continuous *Skipgram with negative sampling* (SGNS) (Mikolov et al., 2013a).

  - Hamilton et al. (2016a) showed the superiority of SGNS over explicit PPMI-based distributional models in semantic shifts analysis, although they noted that low-rank SVD approximations (Bullinaria and Levy, 2007) can perform on par with SGNS, especially on smaller datasets. Since then, the majority of publications in the field started using dense word representations

Tracing Semantic Shifts Distributionally

- **Works employing other distributional approaches to semantic shifts detection**
  - There is a strong vein of research based on dynamic topic modeling (Blei and Lafferty, 2006; Wang and McCallum, 2006), which learns the evolution of topics over time
    - In Wijaya and Yeniterzi (2011), it helped solve a typical digital humanities task of finding traces of real-world events in the texts. Heyer et al. (2016) employed topic analysis to trace the so-called 'context volatility' of words
  - In the political science, topic models are also sometimes used as proxies to social trends developing over time
    - Mueller and Rauh (2017) employed LDA to predict timing of civil wars and armed conflicts. Frermann and Lapata (2016) drew on these ideas to trace diachronic word senses development
  - But most scholars nowadays seem to prefer parametric distributional models, particularly prediction- based embedding algorithms like SGNS, CBOW or GloVe (Pennington et al., 2014)
    - Following their widespread adoption in NLP in general, they have become the dominant representations for the analysis of diachronic semantic shifts as well

Tracing Semantic Shifts Distributionally

- **Comparing vectors across time**
  - It is straightforward to train separate word embedding models using time-specific corpora containing texts from several different periods. But it is not straightforward to compare word vectors across different models
    - It usually does not make sense to directly calculate cosine similarities between embeddings of one and the same word in two different models. Thus, even when trained on the same data, separate learning runs will produce entirely different numerical vectors
    - To alleviate this, Kulkarni et al. (2015) suggested that
      - One should first align the models to fit them in one vector space, using linear transformations preserving general vector space structure. After that, cosine similarities across models become meaningful and can be used as indicators of semantic shifts
      - They also proposed constructing the time series of a word embedding over time (Taylor, 2000).
      - Notably, almost simultaneously the idea of aligning diachronic word embedding models using a distance-preserving projection technique was proposed by Zhang et al. (2015)

Tracing Semantic Shifts Distributionally

- **Eger and Mehler (2016) compared word meaning using 'second-order embeddings'**
  - The vectors of words' similarities to all other words in the shared vocabulary of all models. This approach does not require any transformations: basically, one simply analyzes the word's position compared to other words.
  - At the same time, Hamilton et al. (2016a) and Hamilton et al. (2016c) showed that these two approaches can be used simultaneously: they employed both 'second order embeddings' and orthogonal Procrustes transformations to align diachronic models.
- **It is possible to learn the word embeddings across several time periods jointly, enforcing alignment across all of them simultaneously, and positioning all the models in the same vector space in one step**
  - It was shown in dynamic skip-gram' model and 'dynamic Word2Vec' model. This develops the idea of model alignment even further and eliminates the need to first learn separate embeddings for each time period, and then align subsequent model pairs.
  - A similar approach is taken by Rosenfeld and Erk (2018) who train a deep neural network on word and time representations. Word vectors in this setup turn into linear transformations applied to a continuous time variable, and thus producing an embedding of word *w* at time *t*

Tracing Semantic Shifts Distributionally

- **Another way to make the models comparable**

  - Prediction-based word embedding approaches (as well as RI) allow for incremental updates of the models with new data without any modifications. This is not the case for the traditional explicit count-based algorithms, which usually require a computationally expensive dimensionality reduction step.

  - Kim et al. (2014) proposed the idea of incrementally updated diachronic embedding models:

    - that is, they train a model on the year y_i , and then the model for the year y_{i+1} is initialized with the word vectors from y_i.

    - **This can be considered as an alternative to model alignment**: instead of aligning models trained from scratch on different time periods, one starts with training a model on the diachronically first period, and then updates this same model with the data from the successive time periods, saving its state each time.

    - Thus, all the models are inherently related to each other, which, again, makes it possible to directly calculate cosine similarities between the same word in different time period models, or at least makes the models more comparable

Tracing Semantic Shifts Distributionally

- **Newer research prior to the publication of this paper**
  - Several works aim to address the technical issues accompanying this approach of incremental updating.
    - Among others, Peng et al. (2017) described a novel method of incrementally learning the hierarchical softmax function for the CBOW and Continuous Skipgram algorithms
      - In this way, one can
        - update word embedding models with new data and new vocabulary much more efficiently, achieving faster training than when doing it from scratch, while at the same time preserving comparable performance.
    - Continuing this line of research, Kaji and Kobayashi (2017) proposed a conceptually similar incremental extension for negative sampling
      - It is a method of training examples selection, widely used with prediction-based models as a faster replacement for hierarchical softmax

Tracing Semantic Shifts Distributionally

- **Researchers have to choose the exact method of comparing word vectors across these models**
  - Hamilton et al. (2016a) and Hamilton et al. (2016c) made an important observation that the distinction between linguistic and cultural semantic shifts is correlated with the distinction between global and local embedding comparison methods
    - The former take into account the whole model (for example, 'second-order embeddings,' when we compare the word's similarities to all other words in the lexicon), while the latter focus on the word's immediate neighborhood (for example, when comparing the lists of k nearest neighbors).
  - They concluded that
    - global measures are sensitive to regular processes of linguistic shifts
    - local measures are better suited to detect slight cultural shifts in word meaning
    - Thus, the choice of particular embedding comparison approach should depend on what type of semantic shifts one seeks to detect

Laws of Semantic Change

- **The use of diachronic word embeddings for studying the dynamics of word meaning has resulted in several hypothesized ‘laws’ of semantic change**
  - Law Generalization
    - Dubossarsky et al. (2015) experimented with K-means clustering applied to SGNS embeddings trained for evenly sized yearly samples for the period 1850–2009. They proposed that the likelihood for semantic shift correlates with the degree of prototypicality (the ‘law of prototypicality’ in Dubossarsky et al. (2017)).
    - Another relevant study is reported by Eger and Mehler (2016), based on two different graph models. Based on linear relationships observed in the graphs, Eger and Mehler (2016) postulate two ‘laws’ of semantic change: Experiments were performed with time-indexed historical corpora of English, German and Latin, using time-periods corresponding to decades, years and centuries, respectively.
      - word vectors can be expressed as linear combinations of their neighbors in previous time periods
      - the meaning of words tend to decay linearly in time, in terms of the similarity of a word to itself; this is inline with the ‘law of differentiation’ proposed by Xu and Kemp (2015)

Laws of Semantic Change

- **The use of diachronic word embeddings for studying the dynamics of word meaning has resulted in several hypothesized 'laws' of semantic change**
  - Law Generalization
    - Hamilton et al. (2016a) considered historical corpora for English, German, French and Chinese, spanning 200 years and using time spans of decadesAs in Eger and Mehler (2016), the rate of semantic change was quantified by self-similarity across time-points (with words represented by Procrustes-aligned SVD embeddings). Through a regression analysis, Hamilton et al. (2016a) investi- gated how the change rates correlate with frequency and polysemy, and proposed another two 'laws':
      - 1. frequent words change more slowly ('the law of conformity');
      - 2. polysemous words (controlled for frequency) change more quickly ('the law of innovation').
    - Azarbonyad et al. (2017) showed that these laws (at least the law of conformity) hold not only for diachronic corpora, but also for other 'viewpoints'. However, the temporal dimension allows for a view of the corpora under analysis as a sequence, making the notion of 'semantic shift' more meaningful.

Laws of Semantic Change

- **Introduction to research questioning the validity of patterns of semantic change**
  - Later, Dubossarsky et al. (2017) questioned the validity of some of these proposed 'laws' of semantic change.
    - **Contribution**: In a series of replication and control experiments, they demonstrated that some of the regularities observed in previous studies are largely artifacts of the models used and frequency effects.
    - In particular, they considered 10-year bins comprising equally sized yearly samples from Google Books 5-grams of English fiction for the period 1990–1999
    - For control experiments, they constructed two additional data sets;
      - one with chronologically shuffled data where each bin contains data from all decades evenly distributed,
      - and one synchronous variant containing repeated random samples from the year 1999 alone.
    - Any measured semantic shifts within these two alternative data sets would have to be due to random sampling noise

Laws of Semantic Change

- **Introduction to research questioning the validity of patterns of semantic change**
  - **Conclusion**: Dubossarsky et al. (2017) performed experiments using raw co-occurrence counts, PPMI weighted counts, and SVD transformations (Procrustes aligned), and conclude that the ‘laws’ proposed in previous studies are not valid as they are also observed in the control conditions
    - Semantic change is correlated with frequency, polysemy (Hamilton et al., 2016a) and prototypicality (Dubossarsky et al., 2015)
  - **Why**:
    - Dubossarsky et al. (2017) suggested that these spurious effects are instead due to the type of word representation used – count vectors –
    - and that semantic shifts must be explained by a more diverse set of factors than distributional ones alone.
- **Thus**, the discussion on the existence of the ‘laws of semantic change’ manifested by distributional trends is still open

Diachronic semantic relations

- **Word embedding models are known to successfully capture complex relationships between concepts**
  - As manifested in the well-known word analogies task (Mikolov et al., 2013a), where a model must ‘solve’ equations of the form ‘A is to B is as C is to what?’ A famous example is the distributional model captur- ing the fact that the relation between ‘man’ and ‘woman’ is the same as between ‘king’ and ‘queen ’
    - Thus, it is a natural development to investigate whether changes in semantic relationships across time can also be traced by looking at the diachronic development of distributional models
- **Zhang et al. (2015) considered the temporal correspondences problem**
  - Wherein the objective is to identify the word in a target time period which corresponds to a query term in the source time period (for example, given the query term iPod in the 2000s, the counterpart term in the 1980s time period is Walkman). This is proposed as a means to improve the results of information retrieval from document collections with significant time spans.
    - Szymanski (2017) frames this as the temporal word analogy problem, extending the word analogies concept into the temporal dimension. This work shows that diachronic word embeddings can successfully model relations like ‘word w1 at time period t   is like word w2 at time period tβ ’ .

Diachronic semantic relations

- **A variation of this task was studied in Rosin et al. (2017)**
  - Where the authors learn the relatedness of words over time, answering queries like ‘in which time period were the words Obama and president maximally related’ .
  - This technique can be used for a more efficient user query expansion in general- purpose search engines.
- **Kutuzov et al. (2017a) modeled a different semantic relation:**
  - ‘words w1 and w2 at time period t are in the same semantic relation as words w3 and w4 at time period tβ ’ .
  - To trace the temporal dynamics of these relations, they re-applied linear projections learned on sets of w1 and w2 pairs from the model for the period tn to the model trained on the subsequent time period tn+1.
  - This was used to solve the task of detecting lasting or emerging armed conflicts and the violent groups involved in these conflicts.

Applications

- **Linguistic studies which investigate the how and why of semantic shifts**
  - The first category generally involves corpora with longer time spans
    - since linguistic changes happen at a relatively slow pace.
  - Some examples falling into this category include
    - tracking semantic drift of particular words (Kulkarni et al., 2015) or of word sentiment
    - identifying the breakpoints between epochs (Sagi et al., 2011; Mihalcea and Nastase, 2012)
    - studying the laws of se- mantic change at scale (Hamilton et al., 2016c)
    - finding different words with similar meanings at different points in time (Szymanski, 2017).
  - This has been held up as a good use case of deep learning for research in computational linguistics (Manning, 2015), and there are opportunities for future work applying diachronic word embeddings not only in the field of historical linguistics, but also in related areas like sociolinguistics and digital humanities.

Applications

- **Event detection approaches which mine text data for actionable purposes**
  - The second category involves mining texts for cultural semantic shifts (usually on shorter time spans) indicating real-world events.
  - Examples of this category are
    - temporal information retrieval (Rosin et al., 2017),
    - predicting civil turmoils (Kutuzov et al., 2017b; Mueller and Rauh, 2017),
    - or tracing the popularity of entities using norms of word vectors (Yao et al., 2018).
  - They can potentially be employed to improve user experience in production systems or for policy-making in governmental structures.
- **The near future will see a more diverse landscape of applications for diachronic word embeddings, especially related to the real-time analysis of large-scale news streams**

Open Challenges

- **Challenges in the study of temporal aspects of semantic transformations using distributional modelling (including word embeddings)**
  - The existing methods should be expanded to a wider scope of languages.
    - Hamilton et al. (2016a), Kutuzov and Kuzmenko (2018) and others have started to analyze other languages, but the over- whelming majority of publications still apply only to English corpora. It might be the case that the best methodologies are the same for different languages, but this should be shown empirically.
  - There is a clear need to devise algorithms that work on small datasets, as they are very common in historical linguistics, digital humanities, and similar disciplines.
  - Carefully designed and robust gold standard test sets of semantic shifts (of different kinds) should be created. This is a difficult task in itself, but the experience from synchronic word embeddings evaluation (Hill et al., 2015) and other NLP areas proves that it is possible.
  - There is a need for rigorous formal mathematical models of diachronic embeddings. Arguably, this will follow the vein of research in joint learning across several time spans, started by Bamler and Mandt (2017) and Yao et al. (2018), but other directions are also open.

Open Challenges

- **Challenges in the study of temporal aspects of semantic transformations using distributional modelling (including word embeddings)**
  - Most current studies stop after stating the simple fact that a semantic shift has occurred. However, more detailed analysis of the nature of the shift is needed. This includes:
    - 1. Sub-classification of types of semantic shifts (broadening, narrowing, etc).
    - 2. Identifying the source of a shift (for example, linguistic or extra-linguistic causes).
    - 3. Quantifying the weight of senses acquired over time.
    - 4. Identifying groups of words that shift together in correlated ways.
  - The community around diachronic word embeddings research severely lacks relevant forums, like topical workshops or shared tasks.
    - Diachronic text evaluation tasks like SemEval-2015 are important but not enough, since they focus on identifying the time period when a text was authored, not the process of shifting meanings of a word

Thank you