Francesco Periti[1]

Nina Tahmasebi[2]

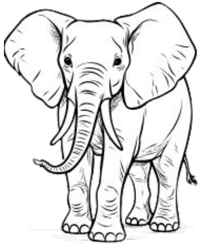
Towards a Complete Solution to Lexical Semantic Change:

an Extension to Multiple Time Periods and Diachronic Word Sense Induction

[1] University of Milan, Italy

[2] University of Gothenburg, Sweden

CHANGE IS KEY.ORG



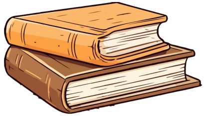
ACL 2024 Bangkok, Thailand

Modeling lexical semantic change through unstructured text

Periti and Montanelli (2024); Tahmasebi et al., (2021)

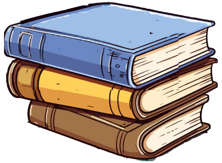
$$\mathcal{C} = \bigcup_{i=1}^{n} C_i$$

C_1



$1700 - 1800$

C_2



$1800 - 1900$

C_3

$1900 - 2024$

detecting
Lexical Semantic Change

w `manufacture`
From *to make by hand*
To *to make by machine*

w `gay`
From *cheerful*
To *homosexual*

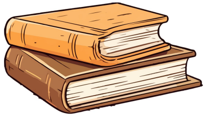
Modeling lexical semantic change through unstructured text

Periti and Montanelli (2024); Tahmasebi et al., (2021)

Simplification

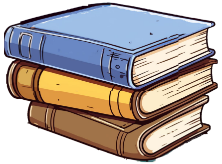
Thus far, the research community has focused on a simplified modeling between two time periods.

C_1



1700 – 1800

C_2



1800 – 1900

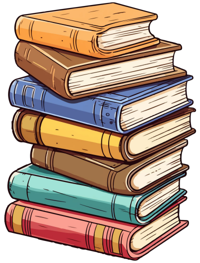
detecting
Lexical Semantic Change

w

`manufacture`
From *to make by hand*
To *to make by machine*

Graded Change Detection

Challenges in handling large corpora



Human annotators

Language models

- Absence of comprehensive benchmarks
- Scalability issues in the computational modeling

Modeling lexical semantic change through unstructured text

Periti and Montanelli (2024); Tahmasebi et al., (2021)

Simplification

Thus far, the research community has focused on a simplified modeling between two time periods.

Modeling semantic change is complex

The simplification serves as a foundational block but is not a *complete* solution. Dealing with hundreds of time periods involves considering millions of comparisons.

$$\langle C_1, C_2 \rangle, \langle C_2, C_3 \rangle, \ldots, \langle C_{n-1}, C_n \rangle$$

$$\langle C_1, C_2 \rangle, \langle C_2, C_3 \rangle, \ldots, \langle C_{n-1}, C_n \rangle$$

$$\langle C_1, C_2 \rangle, \langle C_2, C_3 \rangle, \ldots, \langle C_{n-1}, C_n \rangle$$

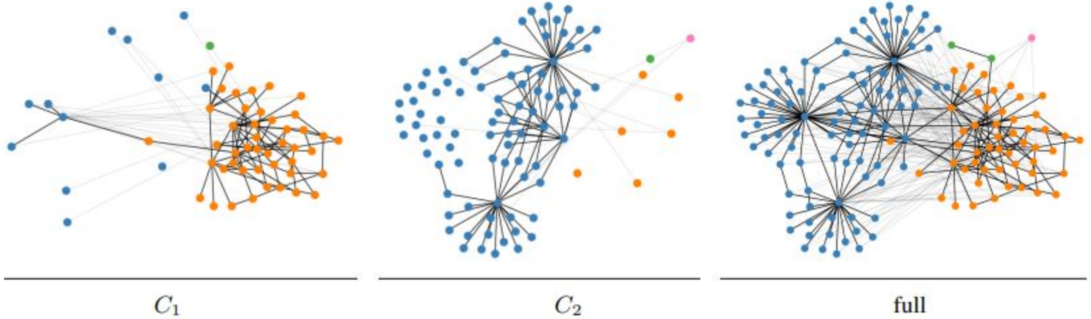


Figure from Laicher et al., (2020)

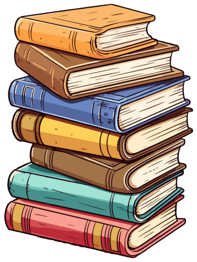
Modeling lexical semantic change through unstructured text

Periti and Montanelli (2024); Tahmasebi et al., (2021)

word usage **extraction** → word occurrence **representation** → word occurrence **aggregation** → semantic change **assessment**

A general workflow for modeling lexical semantic change (Periti and Montanelli, 2024)

target w : plane



Word Sense Induction

form-based

uninterpretable

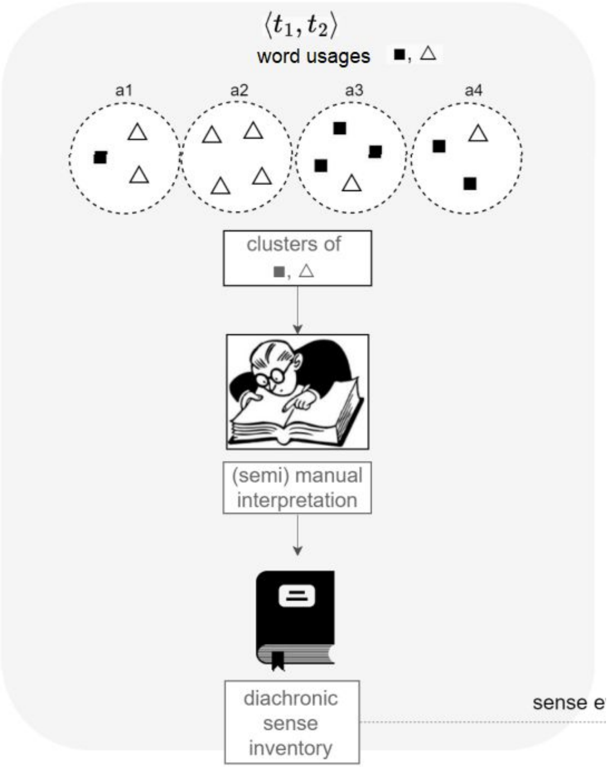
regardless of the number of time periods considered

sense-based

interpretable (?)

Modeling lexical semantic change through unstructured text

[Periti and Montanelli (2024)](); [Tahmasebi et al., (2021)]()



Unsupervised modeling

Modeling word meaning is typically approached in an *unsupervised* manner.

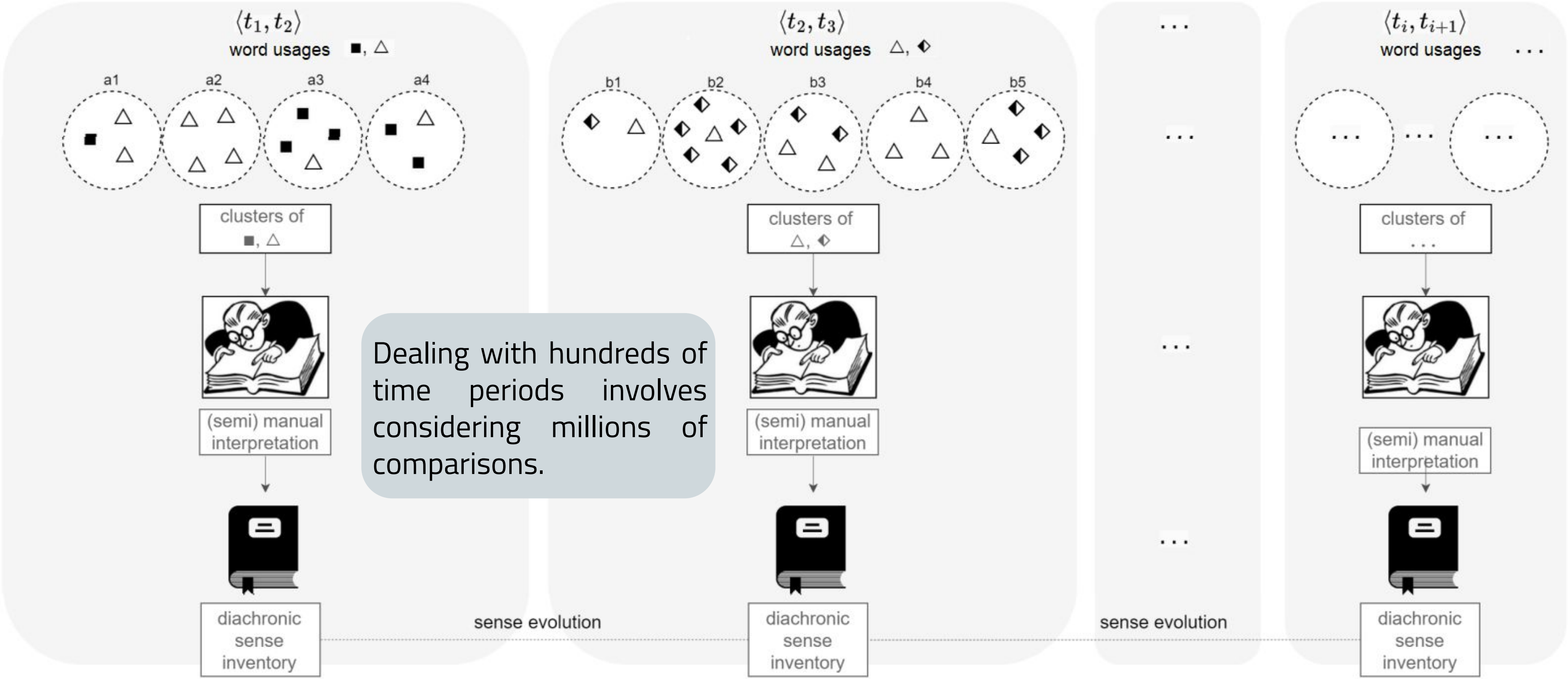
Conventional clustering techniques

Clustering techniques are typically employed to aggregate word usages into *sense* clusters.

Modeling lexical semantic change

Periti and Montanelli (2024); Tahmasebi et al., (2021)

Interpretability
When multiple time periods are considered, challenges increase several orders of magnitude.

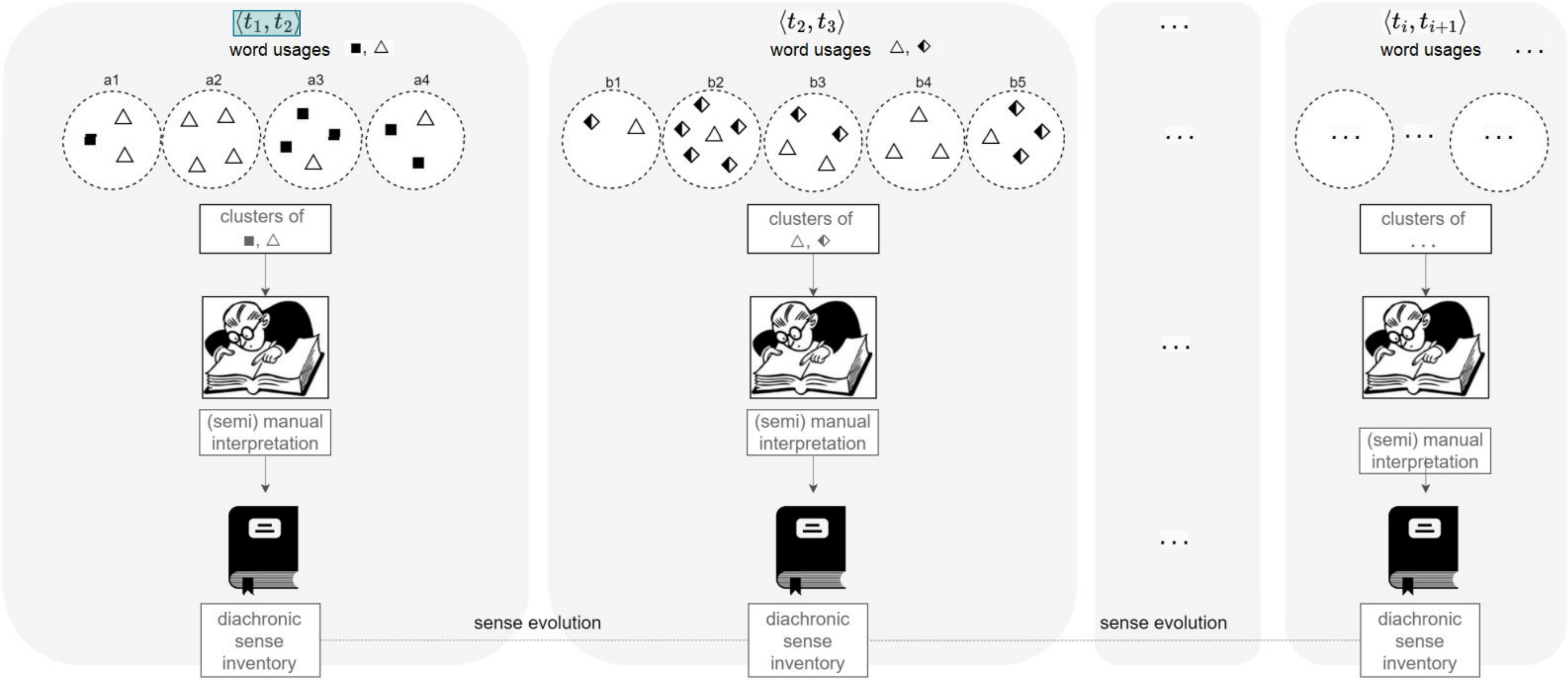


Dealing with hundreds of time periods involves considering millions of comparisons.

Modeling lexical semantic change

Periti and Montanelli (2024); Tahmasebi et al., (2021)

Time
The temporal information of the documents is typically not well-considered.

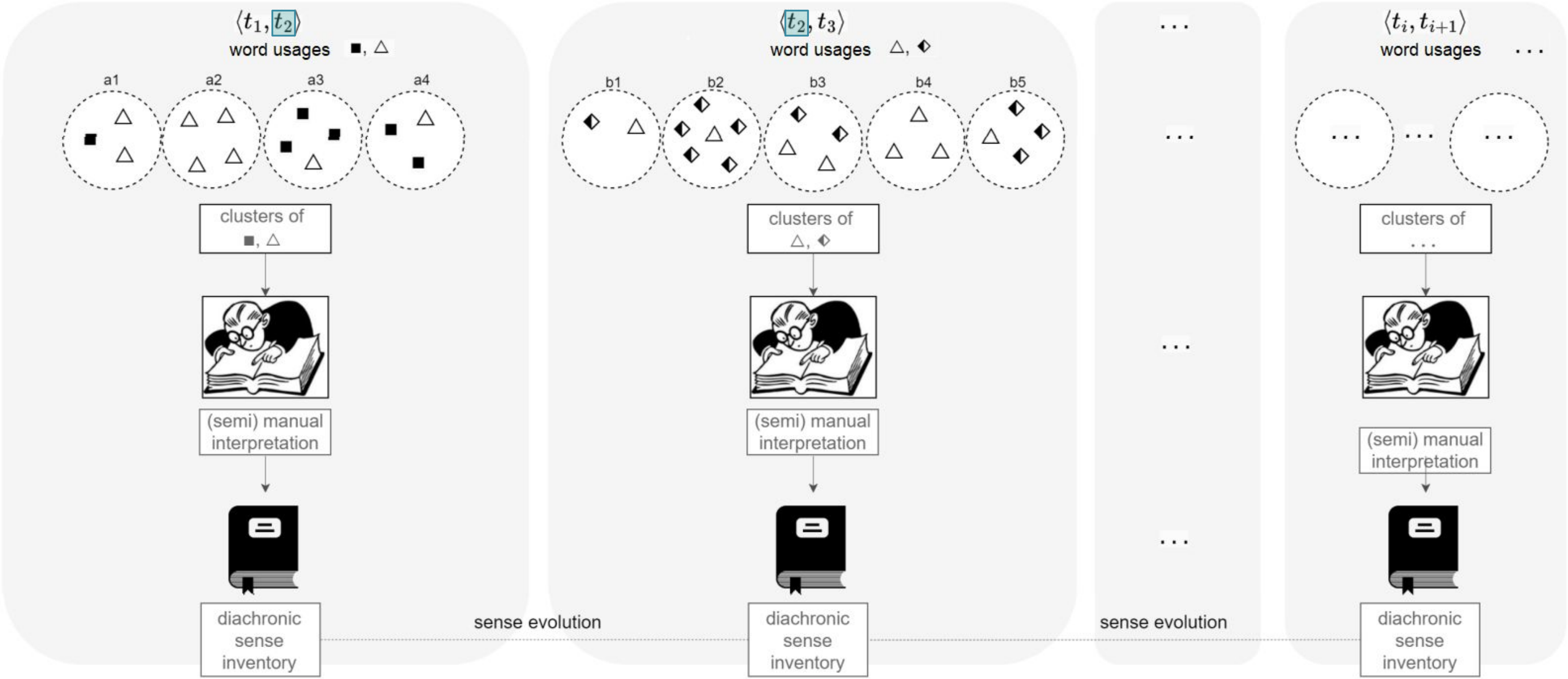


Modeling lexical semantic change

Periti and Montanelli (2024); Tahmasebi et al., (2021)

Inefficiency

Word usages from specific time periods are processed multiple times.



Modeling lexical semantic change through unstructured text

Periti and Montanelli (2024); Tahmasebi et al., (2021)

Extensions to multiple time periods

Diachronic word sense induction

1. Clustering over consecutive time intervals
2. Clustering over consecutive time periods
3. One-time clustering over all time periods
4. Incremental clustering over time periods
5. Scaling up with form-based approaches

Semantic change assessment

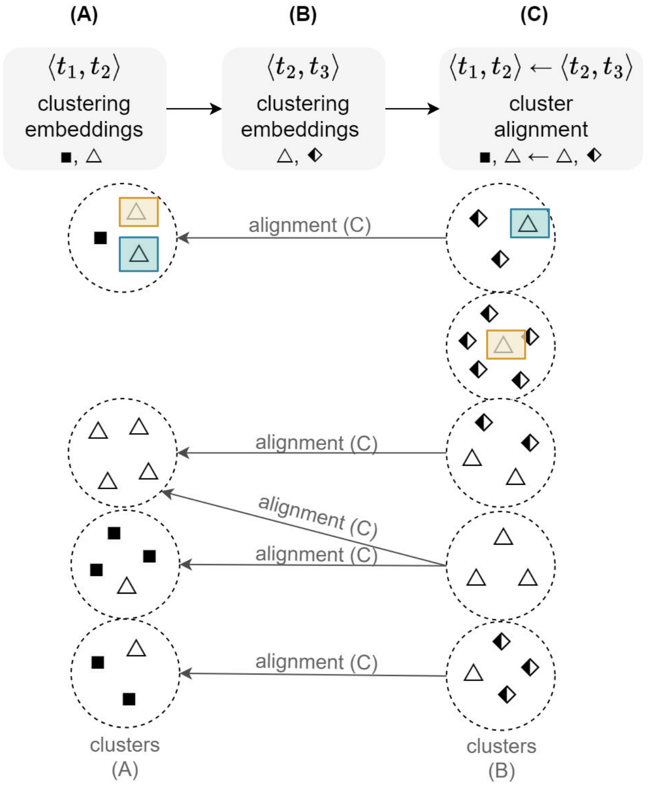
1. Assessment over consecutive time intervals
2. Pairwise assessment over specific time intervals
3. Cumulative assessment over time

There is no best solution that applies to all cases.

Each solution can be implemented with different techniques.

Modeling lexical semantic change over time

1. Clustering over consecutive time intervals



To facilitate the interpretation of sense evolution, a **cluster alignment** step is integrated between consecutive re-executions.

Advantages

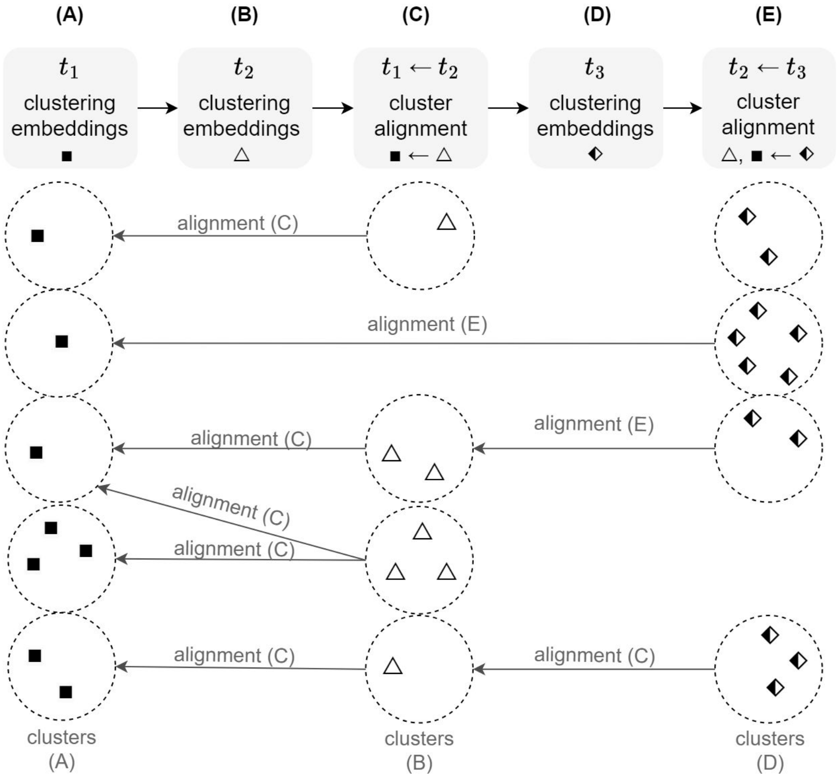
Useful when considering a limited number of word occurrences, as evidence from multiple time periods can enhance the recognition of certain senses.

Disadvantages

Aligning clusters is *complex*.
Clustering is *inefficient* (elements clusters twice).
Clustering arises potential inconsistency.

Modeling lexical semantic change over time

2. Clustering over consecutive time periods



To facilitate the interpretation of sense evolution, a **cluster alignment** step is integrated between consecutive re-executions.

Advantages

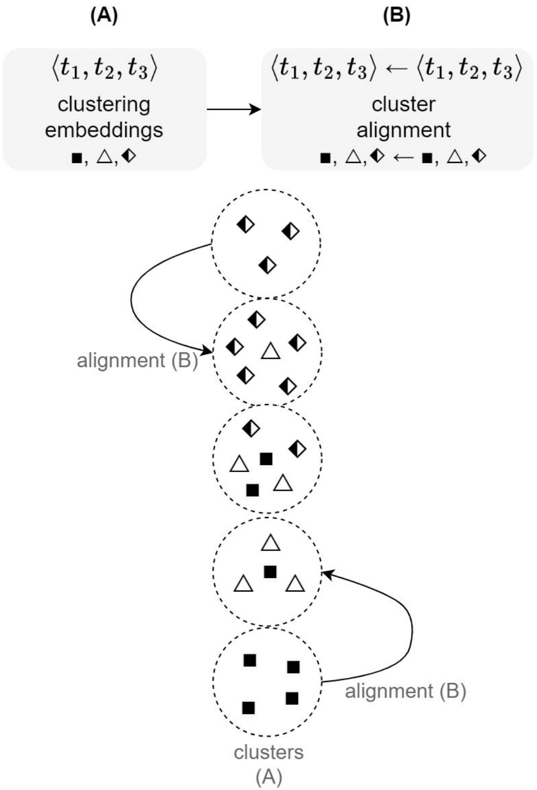
Useful when considering a substantial number of word occurrences.

There is no cluster inconsistency.

Disadvantages

Aligning clusters is *complex*.

Modeling lexical semantic change over time
3. One-time clustering over all time periods



To facilitate the interpretation of sense evolution, a **cluster alignment** step can be **optionally** integrated between consecutive re-executions.

Advantages

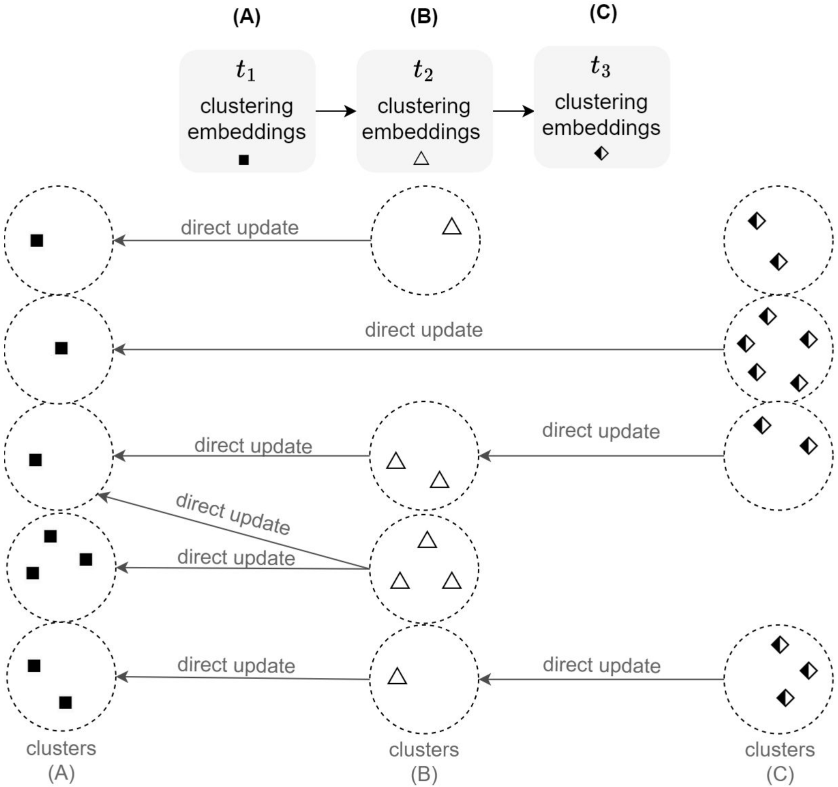
Clustering is simple.
Global view of senses and semantic change.

Disadvantages

Scalability issues may arise.
Unsuitable for data progressively added.
Insensitive to the order of time periods.
No local view of senses and semantic change.

Modeling lexical semantic change over time

4. Incremental clustering over time periods



Incremental clustering algorithms are designed to effectively address the temporal nature of data.

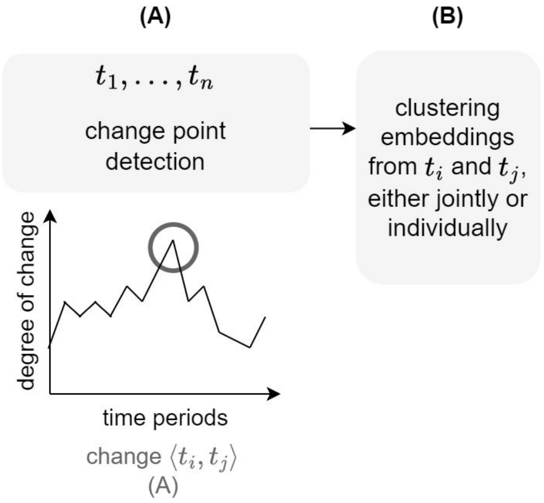
Advantages

Sensitive to the order of time periods.
Suitable for data progressively added.
Clustering alignment is not required.
Evolutionary clustering is flexible.
Evolutionary clustering models sense evolution.

Disadvantages

Evolutionary clustering is *complex*.

Modeling lexical semantic change over time
5. Scaling up with form-based approaches



Most words retain their meaning(s).

Advantages

Scalable.
Simple.
+ advantages of (B)

Disadvantages

A smaller number of word occurrences is considered to model word senses.
+ disadvantages of (B)

Modeling lexical semantic change over time
Semantic change assessments

Assessment over consecutive time intervals

- Presence of noisy artifacts
- Suitable for *immediate change*
- Unsuitable for *periodic senses*

$$\langle t_1,\ t_2\rangle, \langle t_2,\ t_3\rangle, \ldots, \langle t_{n-1},\ t_n\rangle$$

Time series analysis or statistical tests

Pairwise assessment over specific time intervals

- Tailored to specific research questions
- Suitable for *periodic* sense

$$\langle t_{i-1},\ t_i\rangle, \langle t_j,\ t_{j+1}\rangle$$

Cumulative assessment over time

- Novel sense discovery
- Holistic view of semantic change

$$\langle U_{j=1}^{i-1}\ t_j,\ t_i\rangle$$

Towards a Complete Solution to Lexical Semantic Change:
an Extension to Multiple Time Periods and Diachronic Word Sense Induction

Conclusion

We presented possible extensions to expand the modeling over multiple time periods.

Our discussion applies to both the creation of benchmarks and computational modeling.

Optimal modeling must better account for the temporal nature of documents.

Takeaways

The **modeling of word senses is a crucial** for semantic change studies.

There is **no best solution** that applies to all cases: different solutions can be used for different data and purposes.

Each solution can be implemented with **different techniques**.

Optimal modeling must better account for the **temporal nature of documents**.

Francesco Periti[1]

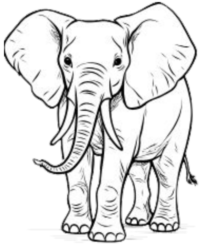
Nina Tahmasebi[2]

Towards a Complete Solution to Lexical Semantic Change: an Extension to Multiple Time Periods and Diachronic Word Sense Induction

[1] University of Milan, Italy

[2] University of Gothenburg, Sweden

CHANGE IS KEY .ORG



ACL 2024 Bangkok, Thailand