Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints

Andrey Kutuzov[1], Mikhail Kopotev[2],
Tatyana Sviridenko[3], Lyubov Ivanova[3]

University of Oslo[1]
University of Helsinki[2]
National Research University Higher School of Economics (Moscow)[3]

Posing the problem

- It is easy to represent documents in monolingual text collections

Posing the problem

- It is easy to represent documents in monolingual text collections
- But suppose the collection contains (comparable) documents in several languages

Posing the problem

- It is easy to represent documents in monolingual text collections
- But suppose the collection contains (comparable) documents in several languages
- Traditional bag-of-words representations fail in this case

Posing the problem

- It is easy to represent documents in monolingual text collections
- But suppose the collection contains (comparable) documents in several languages
- Traditional bag-of-words representations fail in this case
- Human or machine translation can be expensive

Posing the problem

- It is easy to represent documents in monolingual text collections
- But suppose the collection contains (comparable) documents in several languages
- Traditional bag-of-words representations fail in this case
- Human or machine translation can be expensive
- We propose using document representations based on word embeddings

Posing the problem

- It is easy to represent documents in monolingual text collections
- But suppose the collection contains (comparable) documents in several languages
- Traditional bag-of-words representations fail in this case
- Human or machine translation can be expensive
- We propose using document representations based on word embeddings
- Particularly, we compute 'semantic fingerprints' of documents, calculated as average vectors for document words in a given distributional model

Posing the problem

- It is easy to represent documents in monolingual text collections
- But suppose the collection contains (comparable) documents in several languages
- Traditional bag-of-words representations fail in this case
- Human or machine translation can be expensive
- We propose using document representations based on word embeddings
- Particularly, we compute ‘semantic fingerprints’ of documents, calculated as average vectors for document words in a given distributional model
- Then, ‘semantic fingerprints’ from documents in language A are ‘projected’ into language B semantic space, using learned linear transformation matrix.

Description of corpora

We use comparable corpora of Russian and Ukrainian doctoral theses (documents initially collected for the purposes of plagiarism detection).

Description of corpora

We use comparable corpora of Russian and Ukrainian doctoral theses (documents initially collected for the purposes of plagiarism detection).

Corpus	Documents	Word tokens	Contains
Ukrainian	60 000	250 000 000	Extended summaries
Russian	16 000	830 000 000	Full texts

Description of corpora

We use comparable corpora of Russian and Ukrainian doctoral theses (documents initially collected for the purposes of plagiarism detection).

Corpus	Documents	Word tokens	Contains
Ukrainian	60 000	250 000 000	Extended summaries
Russian	16 000	830 000 000	Full texts

Documents are annotated with their topics: economics, history, pedagogy, medicine, law, etc.

Description of corpora

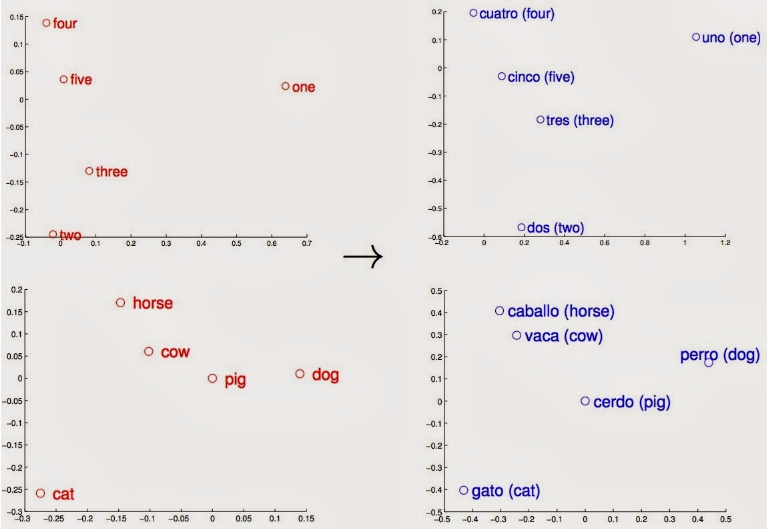
We use comparable corpora of Russian and Ukrainian doctoral theses (documents initially collected for the purposes of plagiarism detection).

Corpus	Documents	Word tokens	Contains
Ukrainian	60 000	250 000 000	Extended summaries
Russian	16 000	830 000 000	Full texts

Documents are annotated with their topics: economics, history, pedagogy, medicine, law, etc.
Now, what kind of document representations can grasp this topical clustering, independent of document language?

Translating models

Semantic structures in distributional models are reproduced even in different languages [Mikolov et al., 2013a]:



Translating models

Continuous Bag-of-Words and Continuous Skip-Gram word embedding models [Mikolov et al., 2013b] were trained on Russian and Ukrainian corpora, with identical hyperparameters:

- vector size 300
- symmetric window of 2 words
- negative sampling (10 samples)
- 5 iterations

This provided us with vector representations of words, such that semantically similar words possess similar vectors.

Translating models

Learning Ukrainian-Russian transformation matrix

1. Semantic spaces of different models (languages) are directly translatable into each other;

Translating models

Learning Ukrainian-Russian transformation matrix

1. Semantic spaces of different models (languages) are directly translatable into each other;
2. Rotate and scale vectors in language **A** in right directions – and you get vectors for their translations in language **B**;

Translating models

Learning Ukrainian-Russian transformation matrix

1. Semantic spaces of different models (languages) are directly translatable into each other;
2. Rotate and scale vectors in language **A** in right directions – and you get vectors for their translations in language **B**;
3. To find right directions, only a small bilingual dictionary is needed, as training data;

Translating models

Learning Ukrainian-Russian transformation matrix

1. Semantic spaces of different models (languages) are directly translatable into each other;
2. Rotate and scale vectors in language **A** in right directions – and you get vectors for their translations in language **B**;
3. To find right directions, only a small bilingual dictionary is needed, as training data;
4. We used about 5 000 Ukrainian-Russian noun pairs

Translating models

Learning Ukrainian-Russian transformation matrix

1. Semantic spaces of different models (languages) are directly translatable into each other;
2. Rotate and scale vectors in language **A** in right directions – and you get vectors for their translations in language **B**;
3. To find right directions, only a small bilingual dictionary is needed, as training data;
4. We used about 5 000 Ukrainian-Russian noun pairs

We have to find the matrix, which is best in transforming the set of 5 000 300-dimensional Ukrainian vectors into the set of their Russian counterparts (also 5000x300).

Translating models

Learning Ukrainian-Russian transformation matrix

1. Semantic spaces of different models (languages) are directly translatable into each other;
2. Rotate and scale vectors in language **A** in right directions – and you get vectors for their translations in language **B**;
3. To find right directions, only a small bilingual dictionary is needed, as training data;
4. We used about 5 000 Ukrainian-Russian noun pairs

We have to find the matrix, which is best in transforming the set of 5 000 300-dimensional Ukrainian vectors into the set of their Russian counterparts (also 5000x300).
In general, such transformation matrix can be computed by solving a linear regression problem: using gradient descent, or numerically (normal equation)

Translating models

Classical normal equation for finding optimal weights for one of 300 target vector components:

$$\boldsymbol{\beta}_i = (\mathbf{X}^\intercal * \mathbf{X})^{-1} * \mathbf{X}^\intercal * y_i \tag{1}$$

X is the matrix of 5000 Ukrainian word vectors (input), y_i is the vector of the i^{th} components of 5000 corresponding Russian words (correct predictions), $\boldsymbol{\beta}_i$ is the vector of 301 optimal coefficients, transforming the Ukrainian vectors into the i^{th} component of the Russian vectors. After solving this for all the 300 target components, we have the full 300x301 transformation matrix.

Translating models

Classical normal equation for finding optimal weights for one of 300 target vector components:

$$\boldsymbol{\beta}_i = (\mathbf{X}^\intercal * \mathbf{X})^{-1} * \mathbf{X}^\intercal * y_i \qquad (1)$$

$\mathbf{X}$ is the matrix of 5000 Ukrainian word vectors (input), y_i is the vector of the i^{th} components of 5000 corresponding Russian words (correct predictions), $\boldsymbol{\beta}_i$ is the vector of 301 optimal coefficients, transforming the Ukrainian vectors into the i^{th} component of the Russian vectors. After solving this for all the 300 target components, we have the full 300x301 transformation matrix.

It is also possible to use regularized normal equation:

$$\boldsymbol{\beta}_i = (\mathbf{X}^\intercal * \mathbf{X} + \lambda * L)^{-1} * \mathbf{X}^\intercal * y_i \qquad (2)$$

Translating models

Classical normal equation for finding optimal weights for one of 300 target vector components:

$$\boldsymbol{\beta}_i = (\mathbf{X}^\intercal * \mathbf{X})^{-1} * \mathbf{X}^\intercal * y_i \tag{1}$$

$\mathbf{X}$ is the matrix of 5000 Ukrainian word vectors (input), y_i is the vector of the i^{th} components of 5000 corresponding Russian words (correct predictions), $\boldsymbol{\beta}_i$ is the vector of 301 optimal coefficients, transforming the Ukrainian vectors into the i^{th} component of the Russian vectors.
After solving this for all the 300 target components, we have the full 300x301 transformation matrix.
It is also possible to use regularized normal equation:

$$\boldsymbol{\beta}_i = (\mathbf{X}^\intercal * \mathbf{X} + \lambda * L)^{-1} * \mathbf{X}^\intercal * y_i \tag{2}$$

...but we found this to be inferior to the non-regularized version.

Inter-lingual translations using vector projections

Now we can predict the corresponding Russian embedding (vector), given the Ukrainian one. Thus, it is possible to 'translate' words.

Inter-lingual translations using vector projections

Now we can predict the corresponding Russian embedding (vector), given the Ukrainian one. Thus, it is possible to 'translate' words.

4500 noun pairs were used as the training set, and 500 noun pairs as the test set. The table shows translation accuracy:

	CBOW		**SkipGram**		**Levenshtein edit distance**
	Training	Test	Training	Test	
@1	0.648	**0.57**	0.545	0.374	0.549
@5	0.764	**0.658**	0.644	0.486	0.619

CBOW is obviously superior to SkipGram in this task.

Inter-lingual translations using vector projections

Now we can predict the corresponding Russian embedding (vector), given the Ukrainian one. Thus, it is possible to 'translate' words.

4500 noun pairs were used as the training set, and 500 noun pairs as the test set. The table shows translation accuracy:

	CBOW		**SkipGram**		**Levenshtein edit distance**
	Training	Test	Training	Test	
@1	0.648	**0.57**	0.545	0.374	0.549
@5	0.764	**0.658**	0.644	0.486	0.619

CBOW is obviously superior to SkipGram in this task.

Examples of translation:

Inter-lingual translations using vector projections

Now we can predict the corresponding Russian embedding (vector), given the Ukrainian one. Thus, it is possible to ‘translate’ words.

4500 noun pairs were used as the training set, and 500 noun pairs as the test set. The table shows translation accuracy:

	CBOW		**SkipGram**		**Levenshtein edit distance**
	Training	Test	Training	Test	
@1	0.648	**0.57**	0.545	0.374	0.549
@5	0.764	**0.658**	0.644	0.486	0.619

CBOW is obviously superior to SkipGram in this task.

Examples of translation:

Ukrainian ‘гетьман’ *hetman* –> Russian ‘царь’ *tzar*,

Inter-lingual translations using vector projections

Now we can predict the corresponding Russian embedding (vector), given the Ukrainian one. Thus, it is possible to 'translate' words.
4500 noun pairs were used as the training set, and 500 noun pairs as the test set. The table shows translation accuracy:

	CBOW		**SkipGram**		**Levenshtein edit distance**
	Training	Test	Training	Test	
@1	0.648	**0.57**	0.545	0.374	0.549
@5	0.764	**0.658**	0.644	0.486	0.619

CBOW is obviously superior to SkipGram in this task.
Examples of translation:
Ukrainian 'гетьман' *hetman* –> Russian 'царь' *tzar*, 'гетман' is the second nearest neighbor.

Inter-lingual translations using vector projections

Now we can predict the corresponding Russian embedding (vector), given the Ukrainian one. Thus, it is possible to 'translate' words.

4500 noun pairs were used as the training set, and 500 noun pairs as the test set. The table shows translation accuracy:

	CBOW		SkipGram		Levenshtein edit distance
	Training	Test	Training	Test	
@1	0.648	**0.57**	0.545	0.374	0.549
@5	0.764	**0.658**	0.644	0.486	0.619

CBOW is obviously superior to SkipGram in this task.

Examples of translation:

Ukrainian 'гетьман' *hetman* –> Russian 'царь' *tzar*, 'гетман' is the second nearest neighbor.

'Nominalization' often occurs, probably because of noun pairs used in the training set:

Ukrainian verb 'розробити' *to develop* –> Russian noun 'разработка' *development*

Semantic fingerprints

Document representations

- Whole documents can be represented as average vectors over all words in them: 'semantic fingerprints'.

Semantic fingerprints

Document representations

- Whole documents can be represented as average vectors over all words in them: 'semantic fingerprints'.
- In our case, they are 300-dimensional (as in the models).

Semantic fingerprints

Document representations

- Whole documents can be represented as average vectors over all words in them: ‘semantic fingerprints’.
- In our case, they are 300-dimensional (as in the models).
- Most interesting: the fingerprints can be projected into another semantic space using the same transformation matrix.

Semantic fingerprints

Document representations

- Whole documents can be represented as average vectors over all words in them: 'semantic fingerprints'.
- In our case, they are 300-dimensional (as in the models).
- Most interesting: the fingerprints can be projected into another semantic space using the same transformation matrix.
- This is what we use to reduce the problem of multi-lingual clustering to the mono-lingual case.

Experimental setting and evaluation

Test set

600 documents were randomly selected.

Language	Economics	Law	History
Russian	100	100	100
Ukrainian	100	100	100

Experimental setting and evaluation

Test set

600 documents were randomly selected.

Language	Economics	Law	History
Russian	100	100	100
Ukrainian	100	100	100

Average document volume is 4 000 word tokens.

Experimental setting and evaluation

Test set

600 documents were randomly selected.

Language	Economics	Law	History
Russian	100	100	100
Ukrainian	100	100	100

Average document volume is 4 000 word tokens.
The aim is to find representations that facilitate clustering documents belonging to one topic into one group, independent of their language.

Experimental setting and evaluation

Evaluation method

Clustering into 3 clusters is performed using the most basic *K-Means* algorithm, to better measure the influence of representations themselves.

Experimental setting and evaluation

Evaluation method

Clustering into 3 clusters is performed using the most basic *K-Means* algorithm, to better measure the influence of representations themselves.

Our evaluation metrics is the accuracy of the resulting clustering in relation to the gold standard division of the test set documents into topics.

Experimental setting and evaluation

Evaluation method

Clustering into 3 clusters is performed using the most basic *K-Means* algorithm, to better measure the influence of representations themselves.

Our evaluation metrics is the accuracy of the resulting clustering in relation to the gold standard division of the test set documents into topics.

E.g., if the majority (more than 100) of **history** documents were assigned to one cluster, we map this cluster to the **history** topic, etc. Then, the ratio of incorrect assignments is calculated.

Experimental setting and evaluation

Evaluation method

Clustering into 3 clusters is performed using the most basic *K-Means* algorithm, to better measure the influence of representations themselves.

Our evaluation metrics is the accuracy of the resulting clustering in relation to the gold standard division of the test set documents into topics.

E.g., if the majority (more than 100) of **history** documents were assigned to one cluster, we map this cluster to the **history** topic, etc. Then, the ratio of incorrect assignments is calculated.

Bag-of-words approach works quite well for monolingual parts of the test set:

Experimental setting and evaluation

Evaluation method

Clustering into 3 clusters is performed using the most basic *K-Means* algorithm, to better measure the influence of representations themselves.

Our evaluation metrics is the accuracy of the resulting clustering in relation to the gold standard division of the test set documents into topics.

E.g., if the majority (more than 100) of **history** documents were assigned to one cluster, we map this cluster to the **history** topic, etc. Then, the ratio of incorrect assignments is calculated.

Bag-of-words approach works quite well for monolingual parts of the test set:

Sub-corpus	Incorrect assignments, %
Ukrainian	4.7
Russian	34.7

Results

Clustering correspondence to the real topics in the multi-lingual setting: 'Binary' and 'count' correspond to bag-of words representations using simple fact of word appearance or using word frequencies

Results

Clustering correspondence to the real topics in the multi-lingual setting: 'Binary' and 'count' correspond to bag-of words representations using simple fact of word appearance or using word frequencies

Method	**Incorrect assignments, %**
Naive Binary	50.17
Naive Count	50.00

Results

Clustering correspondence to the real topics in the multi-lingual setting: 'Binary' and 'count' correspond to bag-of words representations using simple fact of word appearance or using word frequencies

Method	**Incorrect assignments, %**
Naive Binary	50.17
Naive Count	50.00
Edit distance translation Binary	50.50
Edit distance translation Count	50.50

Results

Clustering correspondence to the real topics in the multi-lingual setting: 'Binary' and 'count' correspond to bag-of words representations using simple fact of word appearance or using word frequencies

Method	Incorrect assignments, %
Naive Binary	50.17
Naive Count	50.00
Edit distance translation Binary	50.50
Edit distance translation Count	50.50
Dictionary/Edit distance Binary	50.33
Dictionary/Edit distance Count	49.83

Results

Clustering correspondence to the real topics in the multi-lingual setting: 'Binary' and 'count' correspond to bag-of words representations using simple fact of word appearance or using word frequencies

Method	Incorrect assignments, %
Naive Binary	50.17
Naive Count	50.00
Edit distance translation Binary	50.50
Edit distance translation Count	50.50
Dictionary/Edit distance Binary	50.33
Dictionary/Edit distance Count	49.83
Matrix translation Binary	36.33
Matrix translation Count	36.17

Results

Clustering correspondence to the real topics in the multi-lingual setting: 'Binary' and 'count' correspond to bag-of words representations using simple fact of word appearance or using word frequencies

Method	Incorrect assignments, %
Naive Binary	50.17
Naive Count	50.00
Edit distance translation Binary	50.50
Edit distance translation Count	50.50
Dictionary/Edit distance Binary	50.33
Dictionary/Edit distance Count	49.83
Matrix translation Binary	36.33
Matrix translation Count	36.17
Semantic fingerprints on word types	35.33
Semantic fingerprints on word tokens	**5.50**

Results

Clustering correspondence to the real topics in the multi-lingual setting: 'Binary' and 'count' correspond to bag-of words representations using simple fact of word appearance or using word frequencies

Method	Incorrect assignments, %
Naive Binary	50.17
Naive Count	50.00
Edit distance translation Binary	50.50
Edit distance translation Count	50.50
Dictionary/Edit distance Binary	50.33
Dictionary/Edit distance Count	49.83
Matrix translation Binary	36.33
Matrix translation Count	36.17
Semantic fingerprints on word types	35.33
Semantic fingerprints on word tokens	**5.50**

Results

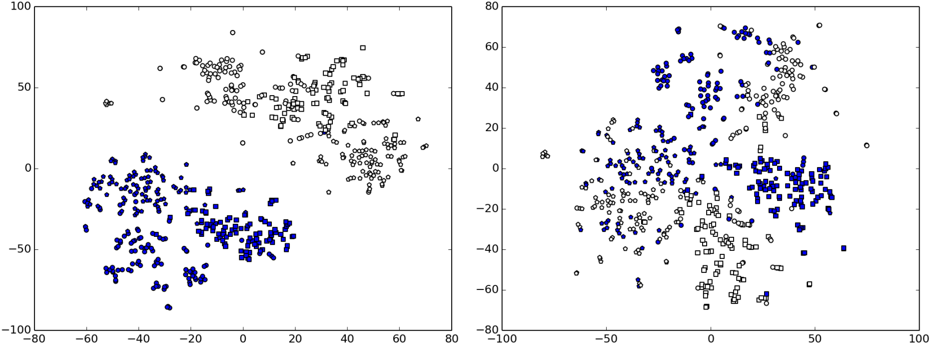
2-dimensional visualizations of document similarities depending on their representations:

colors stand for languages, shapes stand for topics

Results

2-dimensional visualizations of document similarities depending on their representations:

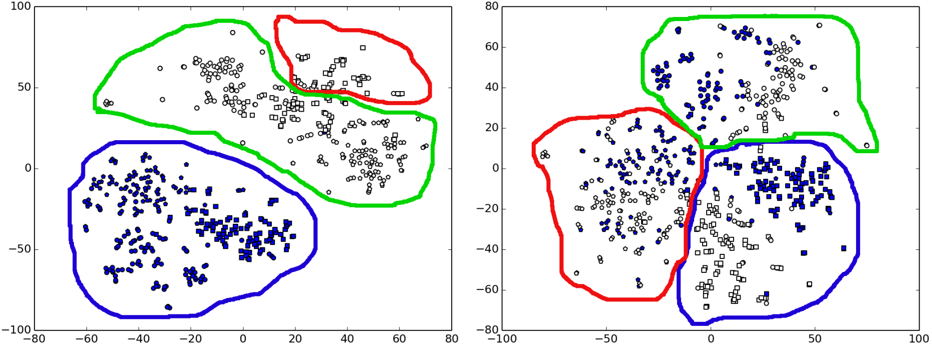
colors stand for languages, shapes stand for topics



Representations with traditional bag-of-words approach (left) and with bag-of-words after ‘translating’ Ukrainian words into Russian using the learned transformation matrix (right).

Results

K-Means clusterings for the collection depending on document representations:



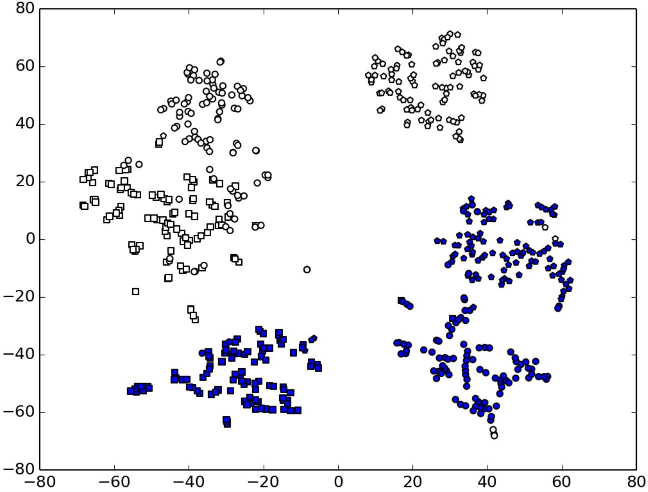
Representations with traditional bag-of-words approach (left) and with bag-of-words after ‘translating’ Ukrainian words into Russian using the learned transformation matrix (right).

Results

2-dimensional visualization of document similarities, when using semantic fingerprints on word tokens:

Results

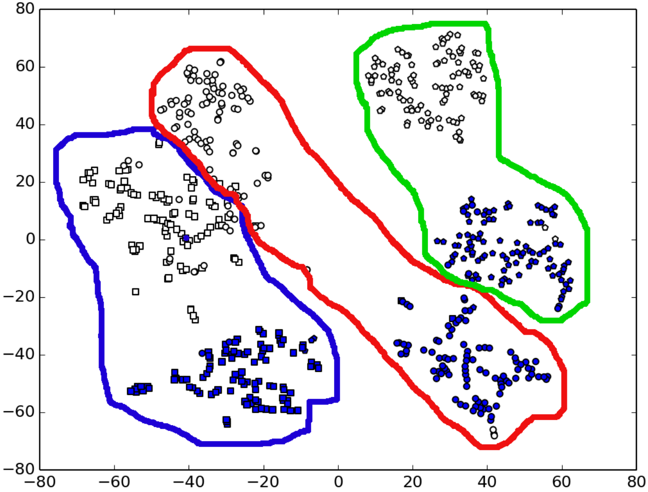
2-dimensional visualization of document similarities, when using semantic fingerprints on word tokens:



(Ukrainian fingerprints projected into Russian with the transformation matrix)

Results

The resulting clustering, when using semantic fingerprints on word tokens:



Discussion

To sum it up:

Discussion

To sum it up:

- Semantic fingerprints are precise enough to reveal topical differences between documents in comparable corpora: F1 measure about 0.95 on our test set, which outperforms [Mathieu et al., 2004]

Discussion

To sum it up:

- Semantic fingerprints are precise enough to reveal topical differences between documents in comparable corpora: F1 measure about 0.95 on our test set, which outperforms [Mathieu et al., 2004]
- This holds even after linear transformation of the fingerprints into another language semantic space.

Discussion

To sum it up:

- Semantic fingerprints are precise enough to reveal topical differences between documents in comparable corpora: F1 measure about 0.95 on our test set, which outperforms [Mathieu et al., 2004]
- This holds even after linear transformation of the fingerprints into another language semantic space.
- Such unified language-independent semantic representations can be used for many cross-lingual NLP tasks

Discussion

To sum it up:

- Semantic fingerprints are precise enough to reveal topical differences between documents in comparable corpora: F1 measure about 0.95 on our test set, which outperforms [Mathieu et al., 2004]
- This holds even after linear transformation of the fingerprints into another language semantic space.
- Such unified language-independent semantic representations can be used for many cross-lingual NLP tasks
- The only resources needed are large comparable monolingual corpora and small bilingual dictionaries (thousands of word pairs).

Discussion

To sum it up:

- Semantic fingerprints are precise enough to reveal topical differences between documents in comparable corpora: F1 measure about 0.95 on our test set, which outperforms [Mathieu et al., 2004]
- This holds even after linear transformation of the fingerprints into another language semantic space.
- Such unified language-independent semantic representations can be used for many cross-lingual NLP tasks
- The only resources needed are large comparable monolingual corpora and small bilingual dictionaries (thousands of word pairs).

Additionally, semantic fingerprints approach is much faster than 'matrix translation'.

Discussion

Why semantic fingerprints are so cool?

Accuracy boost from 65% to 95% is interesting in itself. Why so?

Discussion

Why semantic fingerprints are so cool?

Accuracy boost from 65% to 95% is interesting in itself. Why so?

- Generalized representations do not depend on particular words

Discussion

Why semantic fingerprints are so cool?

Accuracy boost from 65% to 95% is interesting in itself. Why so?

- Generalized representations do not depend on particular words
- With ‘matrix translation’/bag-of words approach even small mistakes lead to dramatic accuracy drop

Discussion

Why semantic fingerprints are so cool?

Accuracy boost from 65% to 95% is interesting in itself. Why so?

- Generalized representations do not depend on particular words
- With ‘matrix translation’/bag-of words approach even small mistakes lead to dramatic accuracy drop
- With semantic fingerprints approach, we take advantage of generalized ‘semantic features’ learned during the model training

Discussion

Why semantic fingerprints are so cool?

Accuracy boost from 65% to 95% is interesting in itself. Why so?

- Generalized representations do not depend on particular words
- With ‘matrix translation’/bag-of words approach even small mistakes lead to dramatic accuracy drop
- With semantic fingerprints approach, we take advantage of generalized ‘semantic features’ learned during the model training
- These features are ‘translated’ way better than particular words

Discussion

Why semantic fingerprints are so cool?

Accuracy boost from 65% to 95% is interesting in itself. Why so?

- Generalized representations do not depend on particular words
- With ‘matrix translation’/bag-of words approach even small mistakes lead to dramatic accuracy drop
- With semantic fingerprints approach, we take advantage of generalized ‘semantic features’ learned during the model training
- These features are ‘translated’ way better than particular words
- Also, in this case topically connected words collectively increase or decrease expression of the corresponding semantic components

Discussion

Why semantic fingerprints are so cool?

Accuracy boost from 65% to 95% is interesting in itself. Why so?

- Generalized representations do not depend on particular words
- With ‘matrix translation’/bag-of words approach even small mistakes lead to dramatic accuracy drop
- With semantic fingerprints approach, we take advantage of generalized ‘semantic features’ learned during the model training
- These features are ‘translated’ way better than particular words
- Also, in this case topically connected words collectively increase or decrease expression of the corresponding semantic components
- Thus, topical words automatically become more important than noise words.

References I

Mathieu, B., Besançon, R., and Fluhr, C. (2004).
Multilingual document clusters discovery.
In *Coupling approaches, coupling media and coupling languages for information retrieval*, pages 116–125.

Mikolov, T., Le, Q., and Sutskever, I. (2013a).
Exploiting similarities among languages for machine translation.
arXiv preprint arXiv:1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b).
Distributed representations of words and phrases and their compositionality.
Advances in Neural Information Processing Systems 26.

Q&A

Clustering Comparable Corpora of Russian and Ukrainian Academic Texts:
Word Embeddings and Semantic Fingerprints

Thank you for your attention!
Questions are welcome.

Andrey Kutuzov, Mikhail Kopotev, Tatyana Sviridenko, Lyubov Ivanova

The trained models, the linear transformation matrix, the evaluation dataset and Python code are available online:
`https://cloud.mail.ru/public/Eune/tN7ssqtWj`