# If Sentences Could See: Investigating Visual Information for Semantic Textual Similarity

Goran Glavas, Ivan Vulic and Simone Paolo Ponzetto (2017)

Robin Ruland, Antonia von Hassell

Ruprecht-Karls-Universität Heidelberg
Institut für Computerlinguistik
Embeddings
Katja Markert, Ines Rehbein
SoSe 2019

9. Juli 2019

# Übersicht

1. Einleitung
2. Multi-modale Repräsentationen
3. Unsupervised STS Maße
4. Evaluation
5. Fazit
6. Kritik
7. Fragen

# Übersicht

1 Einleitung

2 Multi-modale Repräsentationen

3 Unsupervised STS Maße

4 Evaluation

5 Fazit

6 Kritik

7 Fragen

# Einleitung

- betrachtete Task: unsupervised Semantic Textual Similarity (STS)
  - misst Grad an semantischer Äquivalenz zwischen kurzen Texten (i.d.R. Satzpaaren)
- bisherige Ansätze ausschließlich auf linguistischen Modellen basierend

# Einleitung

- Ansatz des Papers: Verwendung von unsupervised multi-modalen Modellen (mit linguistischen und visuellen Informationen) und mehrsprachigen Modellen
- Implementierung der Modelle mit unterschiedlichen Granularitätsebenen:
  - Early fusion (Wortebene)
  - Middle fusion (Satzebene)
  - Late fusion (Fusion der Similarity Scores)

# Übersicht

1 Einleitung

2 Multi-modale Repräsentationen

3 Unsupervised STS Maße

4 Evaluation

5 Fazit

6 Kritik

7 Fragen

# Multi-modale Repräsentationen

- sprachunabhängig
- Mangel an Bildern für ganze Sätze
  - $\Rightarrow$ linguistische und visuelle Repräsentationen für Unigramme (Wörter)
  - Satzrepräsentationen durch Aggregation von Unigrammrepräsentationen

# Linguistische Repräsentationen

- sprachunabhängig (keine sprachspezifischen tools) & Repräsentationen für Unigramme
  - Embeddings
    - englisch: GloVe (Pennington et al., 2014)
    - spanisch, italienisch, kroatisch: Skip-Gram (Mikolov et al., 2013)

# Linguistische Repräsentationen

- für mehrsprachige STS:
  - muss auf den gleichen embedding space projiziert werden
  - mit translation matrix model (Mikolov et al. 2013)

$$\min_{\{S^i, t^i\}_{i=1}^n} \sum_{i=1}^{n} ||s_i * M - t_i||_2$$

  - mithilfe der gelernten Matrix M können dann problemlos Embeddings aus einer Sprache in die andere übersetzt werden, wobei der Informationsverlust für die trainerten Paare minimal ist

# Visuelle Repräsentationen

- n = 20 Bilder pro Wort via Bing

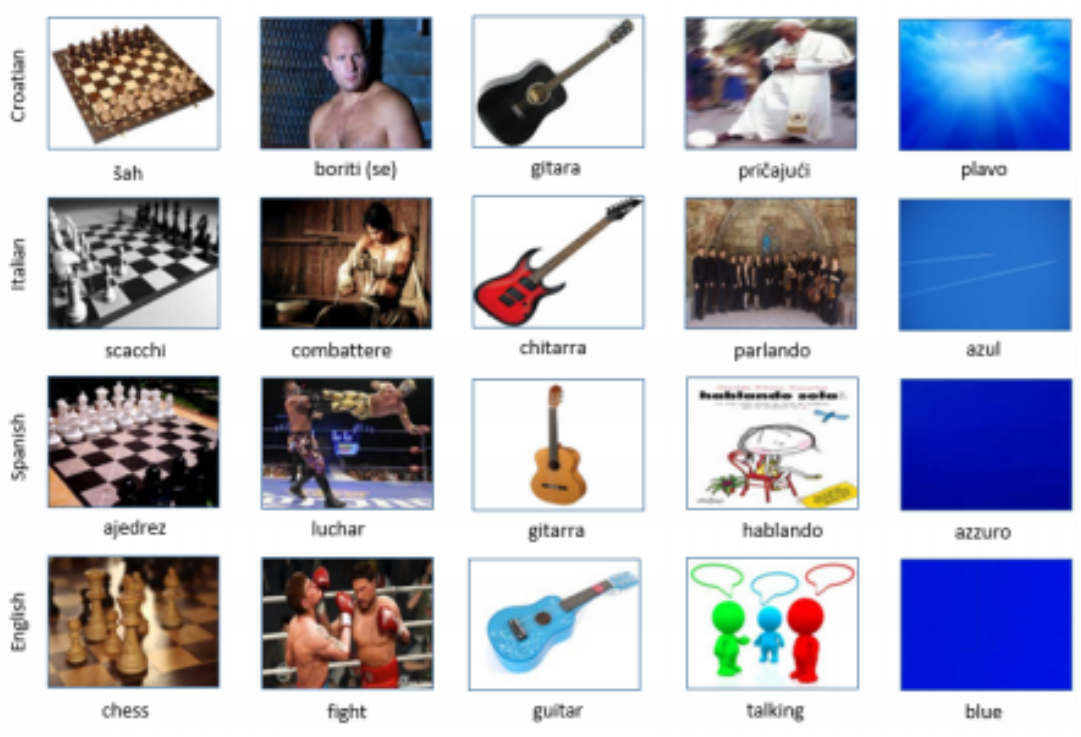
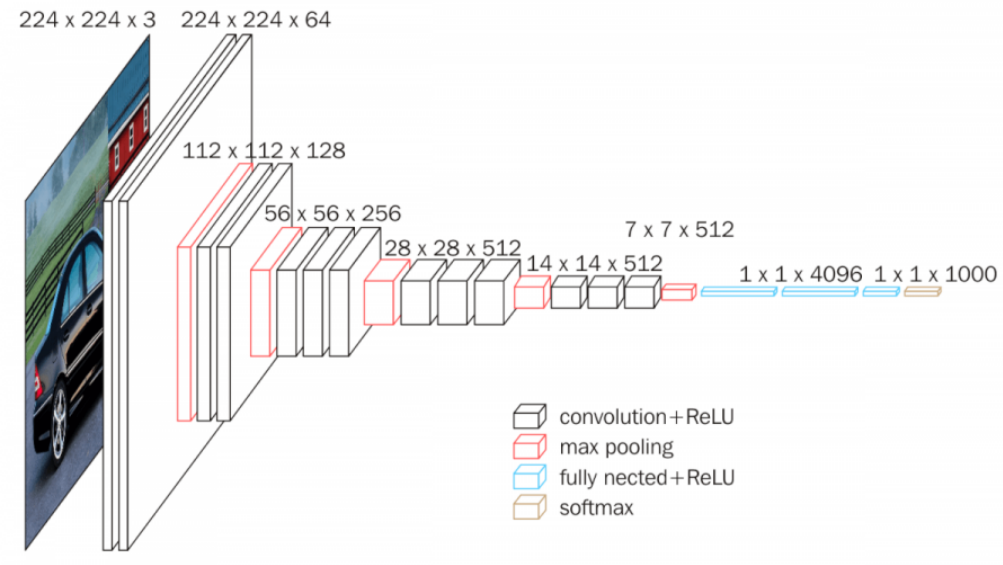



Figure 1: Example images (Bing image search)

# Visuelle Repräsentationen

- deep CNN pre-trained auf dem ImageNet classification task (Russakovsky et al., 2015)




- benutzt pre-softmax Schicht als Embedding
- $\Rightarrow$ visuelle Repräsentation ist Menge von Embeddings

# Multi-modale Repräsentationen

- Early fusion:

$$e_{ef}(w) = e_v(w)||e_t(w)$$

  - $e_v(w)$ ist Durchschnitt oder elementweises Maximum der visuellen Embeddings für ein Wort

- Middle fusion:

$$e_{mf}(S) = (\frac{1}{|S|}\sum_{w \in S} e_v(w))||(\frac{1}{|S|}\sum_{w \in S} e_t(w))$$

# Multi-modale Repräsentationen

- Late fusion:
  - Ähnlichkeit wird getrennt berechnet und gewichtet

$$a * sim_v + b * sim_t$$

  - default ist a = b = 0.5

# Multi-modale Repräsentationen

- selektive Aufnahme
  - semantische Repräsentation verschlechtert sich bei abstrakten Konzepten
  - Idee: Messe Qualität der Bilder und selektiere
  - image dispersion score (Kiela et al., 2014)

$$id(W) = \frac{1}{\binom{|W|}{2}} \sum_{w_i, w_j \in W, i \neq j} 1 - cos(w_i, w_j)$$

  - hoher score bedeutet die Bilder sind verschiedenartig (abstrakte und mehrdeutige Wörter)

# Multi-modale Repräsentationen

- selektive Aufnahme
  - Middle fusion: $\max_{id}(W_1, W_2) > \tau$
  - Late fusion:
    $(1 - \max_{id}(W_1, W_2)) * sim_v + \max_{id}(W_1, W_2) * sim_t$

# Übersicht

1 Einleitung

2 Multi-modale Repräsentationen

3 Unsupervised STS Maße

4 Evaluation

5 Fazit

6 Kritik

7 Fragen

# Unsupervised STS Maße

- optimal aligment similarity
  - zuordnen von Wortpaaren aus den 2 Sätzen:

$$sim_{OA}(S1, S2) = \max_{\{w^i_{S1}, w^i_{S2}\}^N_{i=1}} \sum_{i=1}^{N} sim(w^i_{S1}, w^i_{S2})$$

  - mit Hungarian algorithm (Kuhn, 1955) in polynomialer Zeit
  - normalisiert über die Länge beider Sätze:

$$\frac{1}{2} * (\frac{sim(S1, S2)}{|S1|} + \frac{sim(S1, S2)}{|S2|})$$

# Unsupervised STS Maße

- aggregation similarity
  - berechne Durchschnitt aller Embeddings(Wörter) des Satzes:

$$e(S) = \frac{1}{|S|} \sum_{w \in S} e(w)$$

  - berechne Kosinus-Ähnlichkeit:

$$sim_{agg} = cos(e(S_1), e(S_2))$$

  - Bemerkung: gleich für Early und Middle fusion

# Übersicht

1 Einleitung

2 Multi-modale Repräsentationen

3 Unsupervised STS Maße

4 Evaluation

5 Fazit

6 Kritik

7 Fragen

# Datensätze

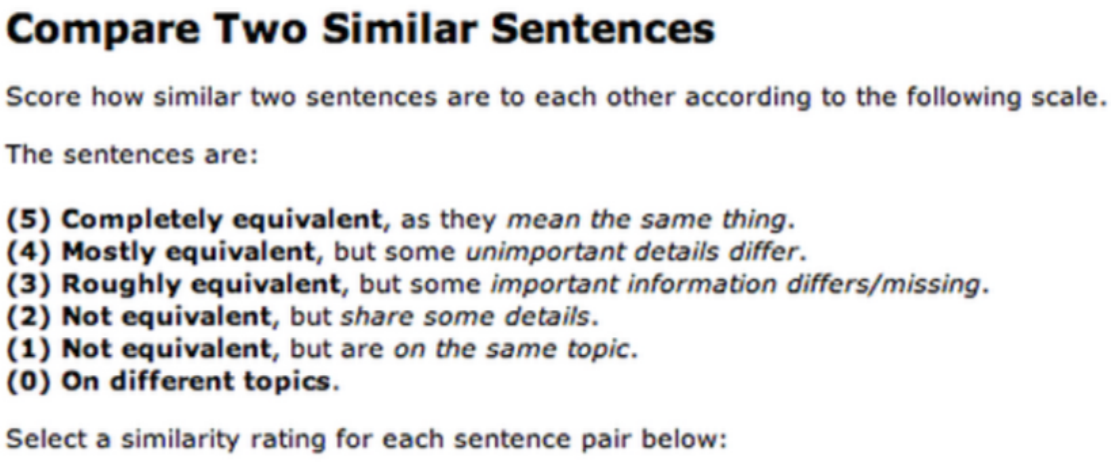
- Verwendung von zwei Datensätzen
- 1. Datensatz: MSRVID
  - Auswertungsteil des Microsoft Research Video Caption Datasets der SemEval 2012 STS challenge (Agirre et al. (2012))
  - enthält 750 Satzpaare mit kurzen englischen Sätzen und eher konkreten Konzepten

# Datensätze: MSRVID


- A person is slicing a cucumber into pieces.
- A chef is slicing a vegetable.
- A person is slicing a cucumber.
- A woman is slicing vegetables.
- A woman is slicing a cucumber.
- A person is slicing cucumber with a knife.
- A person cuts up a piece of cucumber.
- A man is slicing cucumber.
- A man cutting zucchini.

Figure 1: Video and corresponding descriptions from MSRvid

# Datensätze: MSRVID



**Compare Two Similar Sentences**

Score how similar two sentences are to each other according to the following scale.

The sentences are:

**(5) Completely equivalent**, as they *mean the same thing*.
**(4) Mostly equivalent**, but some *unimportant details differ*.
**(3) Roughly equivalent**, but some *important information differs/missing*.
**(2) Not equivalent**, but *share some details*.
**(1) Not equivalent**, but are *on the same topic*.
**(0) On different topics**.

Select a similarity rating for each sentence pair below:

Figure 2: Definition and instructions for annotation

# Datensätze

- Verwendung von zwei Datensätzen
- 2. Datensatz: NEWS-16
  - mehrsprachiger englisch-spanischer STS-Datensatz aus dem SemEval 2016 STS shared task (Agirre et al. (2016))
  - enthält 301 Paare von langen Sätzen aus Nachrichten

# Datensätze: NEWS-16

| Score | English | Cross-lingual Spanish-English |
|---|---|---|
| 5 | *The two sentences are completely equivalent, as they mean the same thing.* | |
| | The bird is bathing in the sink.<br>Birdie is washing itself in the water basin. | El pájaro se esta bañando en el lavabo.<br>Birdie is washing itself in the water basin. |
| 4 | *The two sentences are mostly equivalent, but some unimportant details differ.* | |
| | In May 2010, the troops attempted to invade Kabul.<br>The US army invaded Kabul on May 7th last year, 2010. | En mayo de 2010, las tropas intentaron invadir Kabul.<br>The US army invaded Kabul on May 7th last year, 2010. |
| 3 | *The two sentences are roughly equivalent, but some important information differs/missing.* | |
| | John said he is considered a witness but not a suspect.<br>"He is not a suspect anymore." John said. | John dijo que él es considerado como testigo, y no como sospechoso.<br>"He is not a suspect anymore." John said. |
| 2 | *The two sentences are not equivalent, but share some details.* | |
| | They flew out of the nest in groups.<br>They flew into the nest together. | Ellos volaron del nido en grupos.<br>They flew into the nest together. |
| 1 | *The two sentences are not equivalent, but are on the same topic.* | |
| | The woman is playing the violin.<br>The young lady enjoys listening to the guitar. | La mujer está tocando el violín.<br>The young lady enjoys listening to the guitar. |
| 0 | *The two sentences are completely dissimilar.* | |
| | John went horse back riding at dawn with a whole group of friends.<br>Sunrise at dawn is a magnificent view to take in if you wake up early enough for it. | Al amanecer, Juan se fue a montar a caballo con un grupo de amigos.<br>Sunrise at dawn is a magnificent view to take in if you wake up early enough for it. |

**Table 1:** Similarity scores with explanations and examples for the English and the cross-lingual Spanish-English subtasks.

# Datensätze: Die Sprachvarianten

- Erstellung unterschiedlicher Sprachvarianten
  - MSRVID
    - (EN - EN, EN - ES, EN - IT, EN - HR)
  - NEWS-16
    - (EN - ES, EN - IT, EN - HR)
- Motivation für Sprachwahl:
  - Nutzen vorhandener Ressourcen
  - Verwendung einer Sprache mit wenigen Ressourcen, um Ansatz auf Sprachunabhängigkeit zu überprüfen

# Datensätze

| | MSRVID | | NEWS-16 | |
|---|---|---|---|---|
| Language | ASL | AID | ASL | AID |
| English | 3.42 | 0.51 | 17.0 | 0.68 |
| Spanish | 3.59 | 0.58 | 17.2 | 0.65 |
| Italian | 4.66 | 0.66 | 20.9 | 0.70 |
| Croatian | 3.54 | 0.70 | 17.7 | 0.71 |

Table 2: Statistics of the STS evaluation datasets.

- ASL: average sentence length in number of words
- AID: average image dispersion of words

# Linguistische Embeddings und Translationsmatrizen

- Verwendung trainierter verfügbarer Wortvektoren für
  - Englisch (200-dimensionale GloVe Vektoren trainiert auf 6B Tokens Korpus)
  - Spanisch (300-dimensionale Skip-Gram Vektoren trainiert auf 1.5B Tokens Korpus)
  - Italienisch (300-dimensionale Skip-Gram Vektoren trainiert auf 2B Tokens Korpus)
- Kroatisch (Trainieren von 200-dimensionalen Skip-Gram Embedding Vektoren auf 1.2B Token Version des hrWaC Korpus (Ljubesic und Erjavec, 2011)

# STS Performance

| | MSRVID | | | | NEWS-16 | | |
|---|---|---|---|---|---|---|---|
| Model | EN-EN | EN-ES | EN-IT | EN-HR | EN-ES | EN-IT | EN-HR |
| **Linguistic-only** | | | | | | | |
| TXT-OA | 74.9 | 57.3 | 50.6 | 55.3 | 82.7 | 79.2 | 78.8 |
| TXT-AGG | 74.7 | 54.9 | 42.9 | 51.1 | 57.3 | 48.1 | 54.5 |
| **Visual-only** | | | | | | | |
| VIS-OA-AVG-MAX | 76.5 | 70.4 | 63.1 | 45.0 | 56.9 | 57.5 | 47.7 |
| VIS-AGG-SIM-AVG | 77.6 | 71.8 | 63.1 | 38.2 | 18.0 | 12.4 | 4.4 |
| **Multi-modal** | | | | | | | |
| EF-OA-AVG | 77.0 | 71.5 | 59.7 | 33.3 | 52.8 | 48.9 | 41.7 |
| MF-AVG | 77.8 | 72.0 | 63.8 | 38.9 | 19.1 | 14.8 | 1.3 |
| LF-WORD-OA | 76.6 | 67.9 | 60.9 | 58.3 | 78.1 | 74.9 | 71.0 |
| LF-SENT | 80.8 | **73.1** | **65.4** | 59.2 | 78.0 | 74.0 | 71.3 |
| **Multi-modal with selective inclusion of visual information** | | | | | | | |
| MF-AVG-ID | 78.1 | 70.6 | 50.0 | 53.9 | 57.4 | 50.2 | 54.5 |
| LF-WORD-OA-ID | 77.3 | 64.3 | 56.9 | 58.8 | 82.7 | 79.3 | 78.6 |
| LF-SENT-ID | **81.0** | 71.8 | 63.4 | **61.0** | **83.1** | **79.6** | **79.5** |

Table 3: STS performance on the MSRVID and NEWS-16 datasets (Pearson $\rho$).

# Vergleich mit State-of-the-art

- Vergleich für MSRVID (EN-EN) und EN-ES NEWS-16 (EN-ES) mit besten Systemen der entsprechenden SemEval Tasks
  - Šaric et al. (2012): Pearson Correlation von 88% auf MSRVID; 7% größer als LF-SENT-ID Modell
  - Brychcın und Svoboda (2016): Pearson Correlation von 91% auf EN-ES NEWS -16, 8% größer als LF-SEN-ID

# Übersicht

1 Einleitung

2 Multi-modale Repräsentationen

3 Unsupervised STS Maße

4 Evaluation

5 Fazit

6 Kritik

7 Fragen

# Fazit

- Performanz rein visueller STS Modelle höher als rein linguistischer Modelle bei Datensätzen mit vielen konkreten Konzepten
- multi-modale Modelle erzielen bessere Resultate als uni-modale Modelle
- Performanz rein visueller und multi-modaler Modelle stark abhängig vom Grad der Streuung der Bilder im Datensatz

# Übersicht

1 Einleitung

2 Multi-modale Repräsentationen

3 Unsupervised STS Maße

4 Evaluation

5 Fazit

6 Kritik

7 Fragen

# Kritik

- Normalisierung der Optimal Alignment Similarity
- Warum keine selective inclusion für early fusion?
- Warum bei late fusion nicht auch ein threshold?
- Qulitätsunterschied Aggregation similarity und Optimal aligment similarity
- Qulität der visuellen Embenddings
- selektive Aufnahme: Kosinus-Ähnlichkeit kann auch negativ sein
  - $\Rightarrow$ Problem bei der Formel für selektive Aufnahme

# Übersicht

1 Einleitung

2 Multi-modale Repräsentationen

3 Unsupervised STS Maße

4 Evaluation

5 Fazit

6 Kritik

7 Fragen

# Verständnisfragen

- Abschnitt Einleitung:
  Was sind die Hauptunterschiede zwischen Early Fusion, Middle Fusion und Late Fusion?
- Abschnitt Late Fusion:
  Wie wird sichergestellt, dass die visuellen Signale eine gute Qualität haben?
- Abschnitt Evaluation:
  Warum ist die Performanz für aggregation-based models ähnlich zu entsprechenden optimal-alignment-based models auf dem Datensatz MSRVID, aber deutlich geringer auf dem Datensatz NEWS-16?

# Diskussionsfragen

- Haltet Ihr den Dispersion Score für eine gute Wahl, um die Qualität der Bilder zu beurteilen? Fallen Euch Szenarien ein, in denen informationsreiche Bilder durch den Dispersion Score ausgeschlossen werden könnten?
  - Dispersion Score schließt auch viele Bilder aus, die nicht abstrakt oder polysem sind, z.B. Aktivitäten

Danke für Eure Aufmerksamkeit! Gibt es Fragen?

# Literatur


- Agirre, E., C. Banea, D. Cer, M. Diab, A. Gonzalez-Agirre, R. Mihalcea, G. Rigau, and J. Wiebe (2016). Semeval-2016 Task 1: Semantic textual similarity, mono- lingual and cross-lingual evaluation. In SemEval, pp. 497–511.
- Agirre, E., M. Diab, D. Cer, and A. Gonzalez-Agirre (2012). Semeval-2012 Task 6: A pilot on semantic textual similarity. In SemEval, pp. 385–393.
- Glavaš, Goran, Ivan Vulić, and Simone Paolo Ponzetto. ̈If sentences could see: Investigating visual information for semantic textual similarity.̈IWCS 2017-12th International Conference on Computational Semantics-Long papers. 2017.
- Kiela, D., F. Hill, A. Korhonen, and S. Clark (2014). Improving multi-modal representations using image dispersion: Why less is sometimes mor. In ACL, pp. 835–841.

# Literatur


- Kuhn, H. W. (1955). The hungarian method for the assignment problem. Naval Research Logistics Quarterly 2(1-2), 83–97.
- Ljubesic, N. and T. Erjavec (2011). hrWaC and siWaC: Compiling Web corpora for Croatian and Slovene. In TSD, pp. 395–402.
- Mikolov, T., Q. V. Le, and I. Sutskever (2013). Exploiting similarities among languages for machine translation. CoRR abs/1309.4168.
- Pennington, J., R. Socher, and C. D. Manning (2014). Glove: Global vectors for word representation. In EMNLP, pp. 1532–1543.
- Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpa- thy, A. Khosla, M. Bernstein, et al. (2015). ImageNet large scale visual recogni- tion challenge. International Journal of Computer Vision 115(3), 211–252.