Historical word sense clustering with deep contextualized word embeddings

Bachelor Thesis

Severin Laicher

Prüfer: Apl. Prof. Dr. Sabine Schulte im Walde

Betreuer: Dominik Schlechtweg, Apl. Prof. Dr. Sabine Schulte im Walde

Gliederung

1. Token Vector Representationen
   - Count-Based
   - Word2Vec
   - BERT
2. Tasks
   - Word Sense Clustering
   - Graded LSC Detection
   - Binary LSC Detection
3. Ergebnisse
   - Word Sense Clustering
   - Graded LSC Detection
   - Binary LSC Detection

1. Token Vector Representations

1.1 Count Based

- Automatic Word Sense Discrimination - Schütze
- Zu erst: Type Vektoren erstellen
  - Co-occurrence Matrix
  - Positive Pointwise Mutual Information
  - Singular Value Decomposition
  - 100 Dimensionale Type Vektoren

- Dann: Token Vektoren:
  - Aufsummieren der Type Vektoren im Kontext des Wortes
  - Inverse Document Frequency als Gewicht
  - 100 Dimensionale Token Vektoren

1.2 Word2Vec

- Google's *pre-trained* Word2Vec
- 3 Millionen, 300 dimensionale Type Vektoren
- Basiert auf Skip-Gram Language Modell
- Summier die Type Vektoren aller Kontext Wörter auf

1.3 BERT

- Bidirectional language representation model
- Pretrained Bert-base-uncased model:
- Verarbeitet Text in 12 Layers
  - 12 verschiedene, 768 dimensionale Token Vektoren für jedes Wort
  - Einzeln oder kombiniert
  - Hier: Layer 1 & Layer 12

2. Tasks

2.1 Word Sense Clustering

- Automatic Word Sense Discrimination - Schütze
- Für jede Bedeutung eines ambigen Wortes ein Cluster
- Kmeans
- Group average agglomerative clustering
- Sillhouette Index
- Performance measures:
  - Adjusted Rand Index
  - Cluster Accuracy

1. 137 Pseudowörter mit je 2 Bedeutungen
   - Kmeans mit korrekter Anzahl Cluster
2. 40 SemEval Wörter
   - Anzahl Cluster durch Sillhouette Index

2.2 Graded LSC Detection

- Frage: Wie stark hat sich welches Wort verändert?
- Vergleich mit Gold LSC scores: Spearman correlation coefficient
- Maße (UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection – Kutuzov & Giulianelli):
  - Cosine Similarity
  - Average Pairwise Distance
  - Jensen-Shannon-Distance

2.3 Binary LSC Detection

- Frage: Welche Wörter haben sich verändert?
- Vergleich mit Gold LSC scores: F1
- Maße:
  - Cosine Similarity mit Threshold
  - Average Pairwise Distance mit Threshold
  - Cluster-based (SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection - Schlechtweg et al.):
    - Gibt es ein Cluster, dass zum einen Zeitpunkt mehr als k und zum anderen weniger als l Elemente enthält?

3. Ergebnisse

Pseudowörter:

	BERT	W2V	Count-Based
Mean ARI	0.44	0.27	0.24
Mean ACC	0.80	0.74	0.72

SemEval Wörter

	BERT ohne Lemma	Bert mit Lemma	W2V ohne Lemma	W2V mit Lemma	Count-Based
Mean ARI Kmeans	0.06	0.09	0.03	0.04	0.03
Mean ACC Kmeans	0.61	0.67	0.61	0.64	0.52
Mean ARI AGL	0.08	0.15	0.12	0.13	0.09
Mean ACC AGL	0.64	0.83	0.72	0.78	0.67

1. AGL>Kmeans
2. SemEval – Unbekannte Anzahl Cluster
3. BERT – Position des Wortes
   - Layer 1 & Layer 12
4. BERT – Einfluss der Wortform
5. Reale Verteilung der Wortverwendungen:
   - Durchschnittliches Clustering: [1,1,1,2,6,66]

3.2 Graded LSC Detection

- Übersicht aller (Graded) Spearman Correlation scores

	BERT ohne Lemma	Bert mit Lemma	W2V ohne Lemma	W2V mit Lemma	Count-Based
APD	0.65	0.34	0.36	0.40	0.47
COS	0.40	0.16	0.10	0.18	0.18
JSD Kmeans	0.12	0.11	0.13	0.01	0.01
JSD AGL	0.27	0.20	0.02	0.07	0.08

Vergleich Shared Task:

- Bester Gesamt: 0.527

3.3 Binary LSC Detection

- Übersicht aller (Binary) F1 scores

	BERT ohne Lemma	Bert mit Lemma	W2V ohne Lemma	W2V mit Lemma	Count-Based
APD	0.75	0.44	0.56	0.63	0.56
COS	0.67	0.39	0.56	0.60	0.60
Cluster Based Kmeans	0.56	0.48	0.17	0.30	0.46
Cluster Based AGL	0.56	0.48	0.17	0.29	0.46

Vergleich Shared Task:

- Bester Gesamt: 0.646

- Type>Token?
  - Nur lemmatisiert?
- Bert nicht lemmatisiert am besten

DANKE!