Seminar für Computerlinguistik
Universität Heidelberg
Seminar: Multiword Expression
Dozent: Dr. Mark-Christoph Müller

Multiword noun compound bracketing using Wikipedia

10.12.2019

Nik Lenz

Thematische Einordnung

- MWE Discovery
- MWE Identification

- Noun Compound bracketing

Zu den Autoren

- Caroline Barrière
- Heute Universität Ottawa
- Paper entstand am Centre de Recherche Informatique de Montréal (CRIM)
- Schwerpunkte der Forschung:
  - Entity Detection & Linking
  - Textual Annotation
  - Domain Classification

Zu den Autoren

- Pierre André Ménard
- Am Centre de Recherche Informatique de Montréal (CRIM)
- Schwerpunkte der Forschung:
    - Term Sense Disambiguation
    - Concept Extraction

Warum noun compound bracketing?

- Dreiwortkomposita enthalten nur zwei mögliche Subkomposita.
- Deshalb kann leicht festgelegt werden, welche beiden Worte ein Subkompositum bilden.
- Komposita mit mehr als 3 Worten erfordern andere Herangehensweisen.

Warum noun compound bracketing?

- Lange Nominalkomposita können Dependenzen über große Distanzen enthalten

Beispiel:

wooden French onion soup bowl handle

enthält eine Dependenz zwischen *wooden* und *handle*

Warum noun compound bracketing?

- Bracketing dient der Abbildung der zusammengehörigen Subgruppen eines längeren Kompositums.

  Beispiel:

  (*wooden (((French (onion soup)) bowl) handle))*

Bracketing Algorithmus

- Grundannahme: Dependenzen generell von links nach rechts.

Bracketing Algorithmus

- Grundannahme: Dependenzen generell von links nach rechts.

Bracketing Algorithmus

- Ausdruck:

  wooden soup bowl handle

L1

Wortpaar	Dependency Score
wooden, soup	0,2
wooden, bowl	0,7
wooden, handle	0,5
soup, bowl	0,8
soup, handle	0,1
bowl, handle	0,5

Bracketing Algorithmus

- Ausdruck:

  ((*wooden (soup bowl)) handle)*

L1

Wortpaar	Dependency Score
wooden, soup	0,2
wooden, bowl	0,7
wooden, handle	0,5
soup, bowl	0,8
soup, handle	0,1
bowl, handle	0,5

L2

Wortpaar	Dependency Score
soup, bowl	0,8
wooden, bowl	0,7

- Algorithmus ist greedy, d.h. immer der Eintrag mit dem höchsten Score wird ausgewählt.

Basic dependency association

- Basiert nur auf Kookkurrenz der zwei Wörter.
- Stark korpusabhängig (Größe und Domäne)
- Als association measure hier Dice und Pointwise Mutual Information.

Dice

- Ähnlickeitsmaß für zwei Terme
- Bildet die Häufigkeit von n-Grammen ab, die in beiden Termen vorkommen.
- Hier vermutl. die Häufigkeit von Bigrammen im Korpus.

Pointwise Mutual Information

- Modelliert die Diskrepanz zwischen der Wahrscheinlichkeit der Unabhängigkeit zweier Wörter für ihre gemeinsame und unabhängige Verteilung.
- Je häufiger Wörter zusammen auftreten, desto höher der PMI-Score.

Association Model

- 3 Typen Associations:
  - Relational
  - Koordiniert/Gleichrangig
  - Lexikalisch
- Dienen zur Modulation des Dependency Score des Bracketing Algorithmus.

Modulation Factor I

- Relational association
- Vorkommen einer Präposition zwischen den beiden Wörtern wird gezählt.
- Frequenz boostet den basic dependency association score zwischen den beiden Wörtern

Relational association

- Präpositionen *about, at, by, for, from, in, of, on, to, with*
- Suche nach Mustern „Wort1 *at* Wort“

Modulation Factor II

- Coordinate association
- Konjunktionen zwischen den beiden Wörtern werden gezählt.
- Hohe Frequenz senkt den Basic Dependency Score.

Coordinate association

- Konjunktionen *or, and, nor*
- Suche nach Mustern „Wort1 *and* Wort2“

Modulation Factor III

- Lexical association
- Suche nach Subexpressions
- Zwei Ansätze:
    - Statistisch
    - Vorkommen als Wikipediaeintrag

Lexical association

- Statistical approximation
- Det + Plural als Beweis starker lexikalischer Bindung.

  Bsp: „the Wort1 plural(Wort2)“

  "the cotton shirts"
- Boostet den dependency score zwischen beiden Wörtern.

Vorkommen in Wikipedia

- Zwei Strategien:

- Vorkommen erhöht die dependency scores aller Wörter des Subcompound.

- Gefundene LUs werden zu festen Einschränkungen des bracketing Algorithmus.

Compound segmentation

- LUs → segmentation constraints
- Association scores zwischen LUs anstatt Wortpaaren
- Kleinste Zahl von Entitäten im Compound wird ausgewählt.

Dataset

- Goldstandard basiert auf manuell nachannotierten Penn Treebank Daten.
- Aus techn. Gründen wird für alle NEs Kompositionalität angenommen.

Total Unique Compounds	4749	
3-Word Compounds	2889	60,95%
4-Word Compounds	1270	26,79%
5-Word Compounds	413	8,71%
6+ Word Comps. (max 9)	177	3,72%

Evaluation

- 3 Metriken:

Strict: exakter Goldstandard-Match

Lenient: recall gemessen am Goldstandard

- Binary Tree: (A (B C)) wird als A-C und B-C evaluiert.

- Sub-expression: evaluiert wie viele Sub-expressions aus der Goldstandard-Expression getroffen wurden. Top-level Expression wird nicht betrachtet.

Aus (((A B) C) D)

werden nur die Ausdrücke (A B) und (A B) C) betrachtet.

Evaluation

Gold	Evaluated	Gold elements		Strict	Lenient	
		Subexpression	Binary tree		Subexpression	Binary tree
(a b) c	(a b) c	(a b)	a-b, b-c	100%	100%	100%
(a b) c	a (b c)	(a b)	a-b, b-c	0%	0%	50%
(a b) (c d)	(a b) (c d)	(a b), (c d)	a-b, c-d, b-d	100%	100%	100%
(a b) (c d)	a (b (c d))	(a b), (c d)	a-b, b-d, c-d	0%	50%	66.6%
(((a b) c) d) (e f)	a (b (c (d (e f))))	(a b), (a b c), (a b c d), (e f)	a-b, b-c, c-d, d-f, e-f	0%	25%	40%
			Average:	40%	55%	71.3%

Table 1: Applied examples of evaluation metrics.

Results

Resource	Algorithm	Strict	Lenient
Wikipedia	Dice	55,00%	67,63%
	PMI	56,25%	68,98%
Google Web Ngram	Dice	51,80%	63,90%
	PMI	60,41%	72,47%

Results

Option	Strict	Lenient	Binary
Baseline	0,5500	0,6763	0,8132
Only lexic. assoc.	0,5842	0,7106	0,8321
Only relat. assoc.	0,5854	0,7093	0,8314
Only coord. assoc.	0,5867	0,7110	0,8325

Einfluss der korpusbasierten Statistiken (lexic./relat./coord. association).

Option	Strict	Lenient	Binary
Baseline	0,5500	0,6763	0,8132
Entity-based refinement (uniform distr.)	0,6020	0,7257	0,8408
Entity-based comp.-segment.	0,7316	0,8213	0,8940

Einfluss der entitätsbasierten Ansätze.

Corpus-based improvements

- Alle drei Verfahren kommen alleine jeweils auf sehr ähnliche Ergebnisse.
- Kombination führt zu Verschlechterung.

Entity-based improvements

- Score Modulation nicht besonders effektiv.
- Compound Segmentation erzielt beste Ergebnisse.

Kritik

- Methodik wird unklar beschrieben.
-

Literatur

- Barriére, Caroline & Ménard, Pierre André (2014) *Multiword noun compound bracketing using Wikipedia.*