Supervised Induction of a Parallel Corpus with Sense Information

Julia Konstantinova
Karls-Ruprecht Universität Heidelberg
juli_konstantinova@yahoo.de

Albina Galeeva
Karls-Ruprecht Universität Heidelberg
binaka@web.de

Inna Nickel
Karls-Ruprecht Universität Heidelberg
inna_nickel@web.de

1. Problemstellung und Ziele

•die existierenden parallelen Corpora (JRC-Aquis, Europarl) sind nur für die ressourcenreichen Sprachen vorhanden
• Alternativen:
  •Wikipedia => ein multilingualer Corpus
  •ABER: keine lexiko-semantische Annotation vorhanden
• Automatische Erstellung paralleler Corpora (Englisch+Deutsch) mit der Wortbedeutungsinformation auf Englisch
• Wortbedeutungsinformation => Synonyme, die im gegebenen Kontext verwendet werden können

•Bsp.
  • en. John went to the bank to open a new account
  • de. John ging zur Bank__**depository_financial_institution |bank |banking_concern |banking_company**__ um ein neues Konto zu eröffnen

2. Hyperlinkstruktur Wikipedia

• Wikipediaeinträge sind in mehreren Sprachen vorhanden, können aber nicht als parallele Texte betrachtet werden
• Die Wikipediaseiten zu einem Thema in verschiedenen Sprachen können Übersetzungen voneinander sein. Die meisten Artikel sind aber unabhängig voneinander geschrieben
• Viele Sätze in Wikipediatexten beinhalten Verweise (Links) zu anderen zum Thema relevanten Einträgen in Wikipedia
• Die Links sind manuell von den Verfassern erstellt und sind eindeutig
• Daher ist die Annahme: Sätze, die die Links zur gleichen Einheiten beinhalten sind parallel.

3. BabelNet

• BabelNet ist ein großes multilinguales semantisches Netzwerk, das Wortbedeutungsinformationen für verschiedene Sprachen enthält
• Im Projekt verwendete Daten:
  • Mapping von Wikipedialinks zu WordNet
  • Wortbedeutungsinformationen mittels Sensekeys
  • Sensekeys => eindeutige Identifikationsschlüssel zu den Wikipedialinks, die aus WordNet stammen.

• Nutzung von BableNet zur Übertragung der Bedeutungen

4. Projektvorgehensweise

1. Modul:
  •Erstellung des billingualen Lexikons
  •Extraktion relevanter Korporadaten aus Wikipedia (WikipediaEn: WikipediaDE)
  •Bereinigung der XML- Dateien von XML Mark-up Zeichen
  •Speicherung als Textdokument
2. Modul:
  •Extraktion von Sätzen mit Links
  •Repräsentation der Sätze anhand des bilingualen Lexikons
3. Modul:
  •Parallele Sätze finden
4. Modul:
  •Bestimmung der Wortbedeutungen
  •Übertragung der Wortbedeutung in die deutschen Korpora

5. Ergebnisse

• Fehlerfreie Bedeutungsübertragung anhand des vorhandenen Lexikons (BabelNet)
  EN: Chemistry Actinium shows similar chemical behavior to lanthanum
  DE: Das chemische Verhalten ___DEMEANOR |DEMEANOUR |BEHAVIOR |BEHAVIOUR |CONDUCT |DEPORTMENT___ von Actinium ___ACTINIUM |AC |ATOMIC_NUMBER_89___ hnelt sehr dem Lanthan ___LANTHANUM |LA |ATOMIC_NUMBER_57___

Aufgetretene Probleme:

• Formatprobleme (Sonderzeichen, XML-MarkUp)
  EN: ===Flora and fauna File:Wisent.jpg|thumb|right|A wisent in the Biaowiea Forest Phytogeography|Phytogeographically Poland belongs to the Central European province of the Circumboreal Region within the Boreal Kingdom
  DE: Flora und Fauna Datei:Wisent.jpg|miniatur| Wisent

•Informationsstrukturprobleme:
  EN: In most cases toothaches are caused by problems in the tooth or jaw such as [[Dental caries|cavities]] [[gingivitis|gum disease]] the emergence of [[wisdom teeth]] a marginally cracked tooth infected [[dental pulp]] (necessitating [[endodontic therapy|root canal treatment]] or [[dental extraction|extraction]] of the tooth) jaw disease or exposed [[root canal|tooth root]].
  DE: Ursachen für Zahnschmerzen sind u.a. fehlender Zahnschmelz [[Karies]] und entzündliche Krankheiten wie [[Parodontitis]] aber natürlich auch rein mechanische Verletzungen und Beschädigungen.

• Kulturbezogene inhaltliche Unterschiede
  •Artikel EN “Bikini.txt”
  DE “Bikini.txt”

6. Evaluation

Manuelle Evaluierung auf Basis von 28 Artikeln mit bis zu 580 Sätzen

Titel		Parallele Sätze		
Englisch	**Deutsch**	**gefunden (true positive+false positive)**	**richtig (true positives)**	**nicht gefunden (false negative)**
Cornflower	Kornblume	5	2	0
Moose	Elch	6	2	0
Poland	Polen	23	4	5
Radon	Radon	5	2	0
Bikini	Bikini	3	1	3
Actinium	Actinium	6	3	1
Sugarcane	Zuckerrohr	4	1	0
Amiricium	Americium	6	1	0
Baptism	Baptism	4	1	0
Ukraine	Ukraine	19	3	6
Verb	Verb	2	1	0
Toothache	Zahnschmerzen	1	1	0
Aluminium	Aluminium	17	4	3
Amber	Bernstein	9	1	0
Art	Kunst	10	0	1
Summe:		120	28	19

Recall	Precision	F-Measure
59,5	23,3	33,09

7. References

Luisa Bentivogli, Emanuele Pianta."**Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus**"In *Natural Language Engineering, Special Issue on Parallel Texts*, Volume 11, Issue 03, September 2005, pp. 247-261.

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009). Athens, Greece.

Och, Franz Josef ,Ney, Hermann. 2003. A Systematic Comparison Of Various Statistical Alignment Models. Computational Linguistics, vol.29, num. 1, pp.19-51.

Sisay Fissaha Adafre and M. de Rijke. Finding Similar Sentences across Multiple Languages in Wikipedia. In: EACL 2006 Workshop on New Text, Wikis and Blogs and Other Dynamic Text Sources, April 2006.

Els Lefever and Véronique Hoste (2009), *SemEval-2010 Task 3: Cross-lingual Word Sense Disambiguation*, Proceedings of the Workshop on Semantic Evaluations: Recent achievements and Future Directions (SEW-2009), Boulder, Colorado, pp.82-87.

Roberto Navigli, Simone Paolo Ponzetto. „BabelNet: Bilding a very large Multilingual Semantic Network“ http://aclweb.org/anthology-new/P/P10/P10-1023.pdf

Roberto Navigli, Simone Paolo Ponzetto. „Knowledge-rich Word Sense Disambiguation Rivaling Supervised System“ http://aclweb.org/anthology-new/P/P10/P10-1154.pdf