RuCoCo: Automatische Koreferenzannotationen fu#r Russisch

,
desi@cis.uni-muenchen.de
Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München

,
mikhaylovaa@hotmail.com
Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München

Einführung

Koreferenzresolution beschäftigt sich mit der Aufgabe unterschiedliche sprachliche Ausdrücke, die die gleichen Entitäten im Text beschreiben, automatisch zu identifizieren. Die meisten state-of-the-art Koreferenzresolutionssysteme basieren auf statistischen Verfahren und verwenden große vorannotierte Trainingskorpora (Pradhan et al., 2011). Die Abhängigkeit von Trainingskorpora stellt ein Problem für die Sprachen dar, für die keine Korpora verfügbar sind, die mit Koreferenzinformationen annotiert sind (Recasens et al., 2010; Pradhan et al., 2012; Zhekova, 2013).

Diese Arbeit beschreibt ein Verfahren zur Gewinnung automatischer Koreferenzannotationen durch parallele Korpora für das Russische. Russisch ist eine der Sprachen, für die es noch keine frei verfügbaren Koreferenzannotationen gibt und die daher nicht mit statistischen Koreferenzsystemen bearbeitet werden können. Unser Ziel ist es, Korpora die mehr als einen Zieltext haben zu benutzen (z.B. mehrere Übersetzungen des gleichen Texts/Quelltexts), um die Koreferenzketten von mehreren Zieltexten in den Quelltext zu projizieren. Unser Vorgehen basiert auf der These, dass die Verwendung mehrerer Zieltexte eine bessere Identifikation der Ketten und Grenzen der Mentions (die potentiell koreferenten Phrasen) ermöglicht. Zusätzlich stellen wir die automatisch gewonnenen Koreferenzannotationen (RuCoCo (aus dem Englischen: Russian Coreference Corpus)) für den russischen Originaltext zur freien Verfügung.

Verwandte Arbeiten

Schon oft wurden parallele Korpora für die Gewinnung automatisch generierter Koreferenzannotationen benutzt (Kobdani et al., 2011; Souza and Orašan, 2011; Rahman and Ng, 2012), das Sprachenpaar Russisch-Deutsch wird allerdings wenig bearbeitet (Grishina and Stede, 2015). Darüber hinaus sind parallele Korpora üblicherweise aus einem Quell- und einem Zieltext gebildet (Ganitkevitch et al., 2013; Dolan and Brockett, 2005) und konzentrieren sich hauptsächlich auf die Bearbeitung der englischen Sprache. In Bezug auf das Russische sind uns keine frei verfügbaren Datensätze bekannt, die auch mehrere Zieltexte enthalten. Ein derartiger Korpus für Russisch, der aus dem Roman *Der Idiot* von Dostojewskij und drei Übersetzungen davon besteht, wurde zwar bei der Österreichische Akademie der Wissenschaften entwickelt, er steht aber derzeit nicht zur freien Verfügung (Biber et al., 2002; Dobrovol'skij, 2014).

Abbildung 1: *Interaktives Alignieren in AluDo durch das Suchwort Sonne.*

Aus diesem Grund haben wir einige Ansätze für die Alignierung von Texten für das Sprachenpaar Deutsch-Russisch implementiert (Zhekova et al., 2014), die in das frei verfügbare und interaktive Online - Alignierungstool AluDo (siehe Abbildung 1) integriert wurden. Mit Hilfe von AluDo wurde ein paralleler Korpus erstellt, der aus dem Original des Dostojewskij-Romans *Verbrechen und Strafe* und drei seiner deutschen Übersetzungen (Dostojewskij, 1924; Dostojewskij, 1956; Dostojewskij, 2010) besteht. Einen Teil des Korpus haben wir bereits frei gegeben (Zhekova et al., 2015). Mit Hilfe dieses Korpus beabsichtigen wir, automatische Koreferenzannotationen für den russischen Quelltext zu erzeugen.

Erstellung von RuCoCo

IMSCoref für Deutsch

Zunächst wird der deutsche Teil mit Koreferenzinformationen versehen, um sie anschließend in den russischen Teil der Paralleltexte projizieren zu können. Dafür wird im ersten Schritt das beste frei verfügbare Koreferenzresolutionssystem (laut die CoNLL Evaluationen), IMSCoref (Björkelund and Farkas, 2012), an die deutsche Sprache angepasst.

Datensätze und Annotationen

Wir verwenden die SemEval-Datensätze für Deutsch (Recasens et al., 2010), die in das CoNLL-Format umgewandelt werden. Für die Erzeugung von Syntaxbäumen (ParseBits in CoNLL-Daten), die in den SemEval-Daten fehlen, wird der Stanford Parser (Rafferty and Manning, 2008) eingesetzt. Named Entities sind im deutschen Datensatz in den SemEval-Daten auch nicht verfügbar, daher werden sie automatisch mit Hilfe von Stanford NER (Finkel et al., 2005) extrahiert, und zwar mit dem hgc-Modell für Deutsch (Faruqui and Padó, 2010).

Featureset

Als Baseline-Features werden die Features verwendet, die in IMSCoref für Englisch entwickelt wurden. Das ursprüngliche Featureset für Englisch enthält allerdings Informationen, die in den SemEval-Daten nicht zur Verfügung stehen. Entsprechend werden diese Features ausgelassen (z.B. Speaker-, Genre-Features, usw). Agreement-Features für Deutsch sind zusätzlich integriert worden.

Evaluation

Tabelle 1: Systemevaluation fu#r Englisch (en) und Deutsch (de) mit den original (Orig), reduzierter (Red) und erweiterter (Erw) Featureset.

Um die Güte des für Deutsch adaptierten IMSCoref sicherzustellen, wurde das System für beide Sprachen (Englisch und Deutsch) einem Testlauf unterzogen. Dazu haben wir vier Featuresets getestet: *enOrg* – das Basisfeatureset von IMSCoref in Englisch; *enRed* – das reduzierte Featureset für Englisch (ohne die Features für die keine entsprechenden Annotationen für Deutsch vorhanden sind); *deRed* – das reduzierte Featureset für Deutsch; *enErw* – das erweiterte Featureset für Deutsch. Die Ergebnisse der Evaluation (anhand Identifikation von Mentions (IM), MUC-, B 3 -, und CEAF E - Metriken sowie der offizielle CoNLL-Score) werden in Tabelle 1 präsentiert. Die Zahlen zeigen, dass die Features, die für die deutsche Sprache nicht vorhanden sind, auch für das englische IMSCoref wenig informativ sind. Ihre Nichtberücksichtigung reduziert die Genauigkeit des Systems von 59.29% für *enOrig* auf 58.69% für *enRed* . Damit ist der CoNLL-Score für die deutsche Basisversion des Systems schon sehr hoch und liegt bei 60.55% (*deRed*). Die oben angesprochene Erweiterung der Features für Deutsch bringt eine kleine, aber signifikante Verbesserung von 62.48%.

Projektion der Koreferenzannotationen vom Deutschen ins Russische

Datenaufbereitung

Um das IMSCoref auf die Bearbeitung der deutschen U#bersetzungen der russisch-deutschen Paralleltexte (Zhekova et al., 2015) (aligniert auf Satzebene) vorzubereiten, werden diese auch in das CoNLL-Format transformiert. Zuerst werden die Texte mit dem Stanford-Tokenizer tokenisiert. Der TreeTagger (Schmid, 1994; Schmid, 1995) wird danach für das POS-Tagging eingesetzt. Abschließend werden die Texte der Datenaufbereitung mit GIZA++ (Och and Ney, 2003) auf Wortebene aligniert.

Kettenprojektion

Wir untersuchen drei verschiedene Ansätze für die Übertragung der Annotationen, die im Folgenden beschrieben sind.

Setting 1 Dieser Ansatz folgt Postolache et al. (2006). Für jede deutsche Mention wird eine entsprechende Menge von russischen Tokens extrahiert. Der Kopf der erzeugten Mention ist das russische Token, welches mit dem Kopf der deutschen Mention aligniert wurde. Anfang und Ende einer Mention im Deutschen werden auch direkt durch die Alignierungen im Russischen festgelegt. Falls beide nicht aligniert werden können, wird im Russischen keine entsprechende Mention übertragen.

Setting 2 Bei diesem Setting wird Dependenzinformation im Russischen verwendet, welche mit dem MaltParser (Nivre et al., 2007; Sharoff and Nivre, 2011) erzeugt wird. Dabei wird der lexikalische Kopf einer Mention im Deutschen durch die Wortalignierung ins Russische abgebildet. Daraufhin wird die Länge der Mention durch die Dependenzinformation erzeugt, indem alle Wörter, die in Relation mit dem Kopf stehen, gruppiert werden.

Setting 3 Dieser Ansatz kombiniert Setting 1 und Setting 2, d.h. falls Setting 1 keine Projektion der Mention erzeugen kann, wird Setting 2 eingesetzt.

Evaluation und Fehleranalyse

Tabelle 2: Ergebnisse der Projektion von Koreferenzinformationen vom Deutschen ins Russische

Da keine Gold-Standard-Annotationen vorliegen, werden vorläufig die ersten 30 Sätze des russischen Textes als ein Dokument im Sinne der Koreferenz betrachtet und manuell (von nur einem Annotator) mit Koreferenzketten annotiert.

Die Ergebnisse sind in Tabelle 2 dargestellt, wo alle Übersetzungen (markiert als 1924, 1956, 2010 in der Tabelle) in allen drei Settings (S1, S2 und S3) aufgeführt sind. Die Ergebnisse zeigen, dass durchaus bei allen drei Übersetzungen vergleichbare Zahlen erreicht werden, was innerhalb der drei Settings nicht der Fall ist. Als Erstes zeigt ein Vergleich zwischen der Identifikation der Mentions (IM) für das erweiterte IMSCoref-System im Deutschen (*deErw* in Tabelle 1) und allen drei benutzten Settings und Übersetzungen, dass die Projektion

der Mentions zu einer starken Reduktion der IM-Ergebnisse für Russisch führt. Für alle drei Übersetzungen ist die Projektion in Setting 3 am erfolgreichsten mit F1-Scores zwischen 43.36% und 51.97%. Diese Ergebnisse liegen mit mehr als 20 Prozentpunkten unter dem IM F1-Score für Deutsch. IM ist jedoch eine sehr wichtige Unteraufgabe der Koreferenzresolution, die in hohem Maße den Erfolg der Koreferenzaufgabe beeinflusst (Zhekova, 2013). Als Folge der niedrigen F1-Scores sind die CoNLL-Scores für Russisch wesentlich niedriger im Vergleich zu den Ergebnissen des Systems für Deutsch.

Eine qualitative Analyse zeigt, dass die meisten Fehler durch falsche Wortalignierungen entstehen. Mentions werden nicht gefunden und übertragen, desweiteren haben die projizierten Mentions oft einen falschen Anfang oder ein falsches Ende und tragen so zu den niedrigen Genauigkeiten bei. Oft führt auch die falsche Alignierung zu einer falschen Identifikation des Kopfes im Russischen, wodurch die Ermittlung der Mentionspan im Setting 3 stark beeinflusst wird. Ein Anteil der Fehler ist auch auf Fehlerprojektion basiert – falsche Ketten im deutschen Text werden auch falsch weitergegeben.

Jedoch ist es unser Hauptziel zu untersuchen, ob Korpora, die mehr als einen Zieltext enthalten, hilfreicher für die Projektion sein können als traditionelle Korpora. Dafür haben wir zwei zusätzliche Experimente durchgeführt (jeweils 1924/1956 und 1924/1956/2010 in Tabelle 2) – 1924/1956 fügt die Koreferenzketten aus den beiden Übersetzungen (1924 und 1956) zusammen, während in 1924/1956/2010 die Ketten aus alle drei Übersetzungen zusammengefügt werden (das wird manuell nur für die ersten 30 Sätze des Originalwerks gemacht womit wir entgegen den Testset evaluieren können). Die Ergebnisse zeigen, dass die Benutzung von mehreren Texten sehr hilfreich sein kann, da sich der CoNLL-Score von 20.11% für die Übersetzung 1924 auf 22.41% für 1924/1956/2010 erhöht. Das ist ein sehr positives Ergebnis. Wir vermuten, dass mit einer besseren Alignierung und qualitativ hochwertigeren Dependenzannotation für das Russische, diese Verbesserung noch größer ausfallen würde.

Zusammenfassung

In dieser Arbeit haben wir gezeigt, dass parallele Korpora mit mehr als einem Zieltext für die Gewinnung von automatischen Koreferenzannotationen sehr hilfreich sein können, und dass dadurch Sprachen, die bislang für state-of-the-art Koreferenzsysteme völlig unerreichbar waren, damit bearbeitet werden können. Zusätzlich werden die Koreferenzannotationen für den russischen Originaltext und das manuell annotierte Testset zur freien Verfügung gestellt , womit weitere Pilotstudien für Koreferenzannotation des Russischen möglich gemacht werden. Wir beabsichtigen, das Testset in Zukunft zu erweitern, um adäquatere Ergebnisse bei der quantitativen Analyse der Projektion der Koreferenzinformationen vom Deutschen ins Russische zu erzielen. Außerdem ist ein Verfahren für eine informierte Verbesserung (anhand von Phrasenstruktur, Self-Learning, usw.) von derartigen Annotationen direkt im Zieltext geplant. Die Adaption des IMSCoref-Systems für Deutsch stellen wir auch zur freien Verfügung.

Fußnoten

1. Benutzername und Kennwort werden nach Nachfrage (desi@cis.uni-muenchen.de) vergeben.
2. Die Übersetzung von Swetlana Geier wurde von Fischer-Verlag nur für Forschungszwecke frei gegeben und ist dafür nicht in dem Packet enthalten.

Bibliographie

Biber Hanno / Breiteneder, Evelyn / Dobrovol'skij, Dmitrij (2002): "Corpus-based study of collocations in the AAC", in: Braasch, Anna / Povlsen, Claus (eds.): *Proceedings of the Tenth EURALEX International Congress, Vol. 1.* Center for Sprogtknologi, Kopenhagen.

Björkelund, Anders / Farkas, Richárd (2012): "Data-driven Multilingual Coreference Resolution using Resolver Stacking", in: *Joint Conference on EMNLP and CoNLL - Shared Task* , Jeju Island, Korea. ACL 49–55.

Dobrovol'skij, Dmitrij (2014): "Russkie obrašcenija v parallel'nych korpusach", in: *Die Welt der Slaven* , LIX, 1: 1-21.

Dolan, William B. / Brockett, Chris (2005): "Automatically constructing a corpus of sentential paraphrases", in: *Third International Workshop on Paraphrasing (IWP2005).* Asia Federation of Natural Language Processing.

Dostojewskij, Fjodor M. (1924): *Verbrechen und Strafe* (Übersetzung von Alexander Eliasberg) . Potsdam: Gustav Kiepenheuer Verlag.

Dostojewskij, Fjodor M. (1956): *Schuld und Sühne* (Übersetzung von Hermann Röhl). Berlin: Aufbau Verlag.

Dostojewskij, Fjodor M. (2010): *Verbrechen und Strafe* (Übersetzung von Swetlana Geier). Frankfurt am Mian: Fischer Taschenbuch Verlag.

Faruqui, Manaal / Padó, Sebastian (2010): "Training and evaluating a German named entity recognizer with semantic generalization", in: *Proceedings of KONVENS 2010* , Saarbrücken, Germany.

Finkel, Jenny Rose / Grenager, Trond / Manning, Christopher (2005): "Incorporating non-local information into information extraction systems by gibbs sampling", in: *Proceedings of the 43rd Annual Meeting on ACL* , ACL '05, Stroudsburg, PA, USA. ACL 363–370.

Ganitkevitch, Juri / Van Durme, Benjamin / Callison-Burch, Chris (2013): "PPDB: The paraphrase database", in: *Proceedings of NAACL-HLT* , Atlanta, Georgia, June. ACL.

Grishina, Yulia / Stede, Manfred (2015): "Knowledgelean projection of coreference chains across

languages", in: *Proceedings of the Eight Workshop on Building and Using Comparable Corpora* , Beijing, China. ACL.

Kobdani, Hamidreza / Schütze, Hinrich / Schiehlen, Michael / Kamp, Hans (2011): "Bootstrapping coreference resolution using word associations", in: *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies* , Portland, Oregon, USA, June. ACL 783–792.

Nivre, Joakim / Hall, Johan / Nilsson, Jens / Chanev, Atanas / Eryigit, Gülsen / Kübler, Sandra / Marinov, Svetoslav / Marsi, Erwin (2007): "MaltParser: A Language-Independent System for Data-Driven Dependency Parsing", in: *Natural Language Engineering* 13, 2: 95–135.

Och, Franz Josef / Ney, Hermann (2003): "A systematic comparison of various statistical alignment models", *Computational Linguistics* 29, 1: 19–51.

Postolache, Oana / Cristea, Dan / Orasan, Constantin (2006): "Transferring coreference chains through word alignment", in: *Proceedings of the 5th International Conference on Language Resources and Evaluation.*

Pradhan, Sameer / Ramshaw, Lance / Marcus, Mitchell / Palmer, Martha / Weischedel, Ralph / Xue, Nianwen (2011): "CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes", in: *Proceedings of the CoNLL 2011: Shared Task* , Portland, Oregon, USA. ACL.

Pradhan, Sameer / Moschitti, Alessandro / Xue, Nianwen / Uryupina, Olga / Zhang, Yuchen (2012): "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes", in: *Joint Conference on EMNLP and CoNLL - Shared Task* , Jeju Island, Korea. ACL.

Rafferty, Anna / Manning, Christopher D. (2008): "Parsing three German treebanks: Lexicalized and unlexicalized baselines", in: Proceedings of the Workshop on Parsing German, Columbus, Ohio, June. ACL 40–46.

Rahman, Altaf / Ng, Vincent (2012): "Translation-based projection for multilingual coreference resolution", in: *Proceedings of NACL:HLT 2012* , Montréal. ACL.

Recasens, Marta / Màrquez, Lluís / Sapena, Emili / Martí, M. Antònia / Taulé, Mariona / Hoste, Véronique / Poesio, Massimo / Versley, Yannick (2010): "SemEval-2010 task 1: Coreference resolution in multiple languages", in: *Proceedings of SemEval 2010* , Uppsala, Sweden. ACL.

Schmid, Helmut (1994): "Probabilistic part-of-speech tagging using decision trees", in: *International Conference on New Methods in Language Processing* , Manchester, UK.

Schmid, Helmut (1995): "Improvements in Part-of-Speech Tagging with an Application to German", in: *Workshop of the Special Interest Group for Linguistic Data and Corpus-based Approaches to Natural Language Processing (EACL 1995 SIGDAT-Workshop)* 47–50.

Sharoff, Serge / Nivre, Joakim (2011): "The proper place of men and machines in language technology. Processing russian without any linguistic knowledge", in: *Proceedings of the International Conference on Computational Linguistics and Artificial Intelligence Dialog 2011*, Moscow.

Souza, Jose Guilherme Camargo / Orasan, Constantin (2011): "Can projected chains in parallel corpora help coreference resolution?", in: Hendrickx, Iris / Devi, Sobha Lalitha / Branco, Antonio / Mitkov, Ruslan (eds.): *Anaphora Processing and Applications* (= Lecture Notes in Computer Science 7099). Berlin / Heidelberg: Springer.

Zhekova, Desislava (2013) *Towards Multilingual Coreference Resolution* . Ph.D. thesis, University of Bremen.

Zhekova, Desislava / Zangenfeind, Robert / Mikhaylova, Alena / Nikolaienko, Tetiana (2014): "Alignment of Multiple Translations for Linguistic Analysis", in: *Proceedings of the The 3rd Annual International Conference on Language, Literature and Linguistics (L3)* , Bangkok, Thailand.

Zhekova, Desislava / Zangenfeind, Robert / Mikhaylova, Alena / Nikolaienko, Tetiana (2015): "Sentence-Alignment and Application of Russian-German Multi-Target Parallel Corpora for Linguistic Analysis and Literary Studies", in: Portela, Manuel (ed.): *Digital Literary Studies* 3 (to appear).