HiER 2013

Griesbaum, Heuwing, Ruppenhofer, Werner (Hrsg.)

HiER 2013

Proceedings des 8. Hildesheimer Evaluierungs- und Retrievalworkshop

Hildesheim, 25.–26. April 2013

J. Griesbaum, B. Heuwing, J. Ruppenhofer, K. Werner (Hrsg.):
HIER 2013 - Proceedings des 8. Hildesheimer Evaluierungs-
und Retrievalworkshop, Hildesheim 2013

Einfluss morphologischer Werkzeuge auf die Information Retrieval-Qualität

am Beispiel des Porter-Stemmers und des Lemmatizers *wmtrans*

Melanie Dick

Universität Hildesheim
melaniedick@gmx.net

Zusammenfassung

Der Lemmatizer *wmtrans* der Canoo Engineering AG (Canoo 2012) und der Porter-Stemmer (Porter 1997) werden hier verglichen. Ziel ist es, jeweils den Einfluss auf die Information Retrieval-Qualität zu analysieren. Basierend auf den Testdaten der CLEF Studie von 2002 wird der Schwerpunkt auf die Analyse einzelner Informationsbedürfnisse gelegt.

Abstract

The lemmatizer *wmtrans* developed by Canoo Engineering AG (Canoo 2012) and the Porter stemmer (Porter 1997) will be compared regarding their influence on information retrieval quality. Based on test data from CLEF 2002.

Einleitung

In der Studie wurde der Lemmatizer der Firma Canoo Engineering AG (Canoo 2012) in Basel, welche diese Studie initiierte, mit dem Porter-Stemmer (Porter 1997) verglichen. Im Kontext des Information Retrieval sollte untersucht werden, ob und wie morphologische Werkzeuge die Treffergenauigkeit sowie die Vollständigkeit in Suchlösungen verbessern. Während der Lemmatizer auf lexikographischem Wissen basiert (also auf Wissen

über viele einzelne Wörter), führt ein Stemmer den Wortabgleich regelbasiert durch.

Methodik

Die CLEF-Studie von 2002 (CLEF 2003) wurde als Testdatensatz verwendet. Aus den 50 Informationsbedürfnissen dieser Studie wurden nach zuvor definierten Regeln ad-hoc-Anfragen generiert. Informationsbedürfnis 123 lautete zum Beispiel *„Heirat Jackson-Presley. Suche Dokumente, die über die vermutete Ehe von Michael Jackson mit Lisa Marie Presley oder ihre Trennung berichten".* Hier wurden die sechs Begriffe *Heirat, Michael, Jackson, Lisa, Presley* und *Trennung* gewählt.

Komposita wie beispielsweise der Begriff *Literaturnobelpreisträger* von Informationsbedürfnis 94 wurden nach vier Varianten zerlegt. In zwei Varianten orientierte sich die Kompositazerlegung jeweils an dem Wissen der Wörterbücher canoo.net und duden.de. Die Version von duden.de wurde angelegt, um dem Aspekt, dass wmtrans durch die Wahl des eigenen Wörterbuchs einen Vorteil erhalten könnte zu entkräften. Der wmtrans-Lemmatizer basiert auf dem canoo.net-Wörterbuch. In einer dritten Version fand keine Wortzerlegung statt. In der Vierten wurden alle Kompositabildungen zerlegt. Der Suchbegriff *Literaturnobelpreisträger* wurde nach dem canoo.net-Wörterbuch in *Literaturnobelpreis* und *Träger* zerlegt. Duden.de kannte den Begriff. In dieser Version sowie in der Version ohne Kompositazerlegung wurde *Literaturnobelpreisträger* verwendet. In der letzten Variante wurde der Suchbegriff in *Literatur, Nobel, Preis* und *Träger* zerlegt.

Die Annäherung an eine authentische Nutzeranfrage war ein weiteres Kriterium. Dies wurde zum Beispiel bei der Anzahl der Suchterme berücksichtigt.

Es wurden fünf morphologische Werkzeuge getestet. Als Baseline wurde eine Suchlösung, welche über kein morphologisches Wissen verfügt, implementiert. Sie führt lediglich einen Abgleich der Schreibweise durch.

Ein weiteres Werkzeug war der regelbasierte Porter-Stemmer. Er basiert auf dem von Martin Porter entwickelten Stemming-Algorithmus (Porter 1997).

Der auf lexikalischem Wissen basierende Lemmatizer wmtrans, welcher auf das morphologische Wörterbuch canoo.net zurückgreift, wurde neben der Basisversion in zwei zusätzlichen Versionen durch Wortzerlegung (also Wortbildungsanalyse, zum Beispiel *Mitglied(er) – staat(en))* erweitert. Eine der zusätzlichen Version erlaubte die Wortzerlegung in der Suchanfrage und den Testdatensätzen, in der zweiten Version wurde die Wortzerlegung nur in den Testdaten und nicht in den Suchtermen eingesetzt.

Die fünf Suchwerkzeuge waren die sogenannte Baseline, der Porter-Stemmer, der wmtrans-Lemmatizer ohne die Fähigkeit der Wortzerlegung, ein wmtrans-Lemmatizer mit der Fähigkeit der Wortzerlegung in den Testdaten und ein wmtrans-Lemmatizer mit der Fähigkeit der Wortzerlegung in den Testdaten und Suchtermen.

Analyse

In der Gesamtübersicht aller Anfragen waren zunächst nur geringe Unterschiede zwischen den Suchwerkzeugen zu erkennen.

Die Recall-Werte der 50 Einzelanfragen aus CLEF variieren stark. Einzelanfragen mit Recall-Werten von 0 bis 1 sind enthalten. Interessant war neben der Gesamtübersicht deshalb vor allem die Analyse der einzelnen Informationsbedürfnisse. Bei Ergebnissen mit einem unterschiedlich hohen Recall zwischen Porter-Stemmer und *wmtrans*-Lemmatizer lassen sich die Unterschiede, die in der differenzierten morphologischen Verarbeitung begründet sind, schnell erkennen.

Anfragen mit einem identischen Recall wurden näher analysiert. Die Precision und insbesondere die Precision nach einer bestimmten Anzahl gefundener Dokumente, können ein Indikator dafür sein, dass eine genauere Untersuchung der morphologischen Verarbeitung an einer Stelle der Ergebnismenge lohnenswert ist. Vor- und Nachteile der beiden morphologischen Werkzeuge und ihre Auswirkung auf das Ranking eines Dokuments in der Ergebnismenge werden in der Evaluation behandelt. Anhand ausgewählter Querys werden verschiedene Suchterme und ihre unterschiedliche morphologische Verarbeitung erörtert. Nur von einer Suchlösung gefundene Suchbegriffe wirken sich auf das Ranking der Ergebnisse aus. Dies kann zu einer Verbesserung, aber auch zur Verschlechterung der Precision führen. In einem weiteren Schritt wurde der Einfluss der maschinellen Wortzerlegung, welche

in zwei Versionen des Lemmatizers integriert wurde, auf das Suchergebnis analysiert.

Beispiel: Informationsbedürfnis 101

Die exemplarische Erörterung an dem Informationsbedürfnis Nummer 101 aus der CLEF-Studie illustriert dies. Aus dem Informationsbedürfnis „*Zypern und die EU. Kann Zypern Mitglied der EU werden?2 w*urden die drei Suchbegriffe *Zypern, EU* und *Mitglied* extrahiert. Interessant ist hier vor allem die Verarbeitung der Begriffe *Mitglied* und *Zypern*. *Zypern* wird im Suchprozess als obligatorischer Begriff verwendet. Dies bedeutet, dass *Zypern* in einem Dokument enthalten sein muss, damit es als relevant von den morphologischen Werkzeugen kategorisiert werden darf und in die Ergebnismenge aufgenommen wird.

Die Baseline konnte lediglich einen Abgleich der identischen Schreibweise durchführen und deshalb nur genau die in der Suchanfrage verwendete Schreibweise, *Zypern* beziehungsweise *Mitglied,* erkennen. Der Recall ist bei der Baseline nicht vollständig.

Der Porter-Stemmer und der *wmtrans*-Lemmatizer konnte mit dem Suchbegriff *Mitglied* auch das Wort *Mitglieder* abgleichen. Die beiden um die Fähigkeit der Kompositazerlegung erweiterten Versionen waren auch in der Lage das Kompositum *Mitgliedsstaaten* zu erkennen. Die Version des *wmtrans*-Lemmatizers, welche um die Fähigkeit der Wortzerlegung in den Suchbegriffen erweitert wurde, zerlegt den Begriff *Mitglieder* in *Mit* und *glied*. *Mit* wurde allerdings als Stoppwort erkannt und deshalb im weiteren Suchprozess ignoriert.

Der Suchbegriff *Zypern* wurde vom Porter-Stemmer auf den Wortstamm *zyp-* reduziert und als Verb behandelt. Diese Wortverarbeitung erlaubt es dem Porter-Stemmer nicht mehr *Zyperns,* den Genitiv des Wortes, mit dem Suchbegriff *Zypern* abzugleichen, was dazu führt, dass fünf relevante Dokumente vom Porter-Stemmer und der Baseline nicht gefunden wurden.

Der Lemmatizer wmtrans konnte den Genitiv *Zyperns* und den Suchbegriff *Zypern* einander zuordnen. Auch das Adjektiv *zyprisch* konnte *Zypern* zugeordnet werden.

Die beiden um die Wortzerlegung erweiterten Versionen des Lemmatizers wmtrans konnten auch *Nordzypern* mit *Zypern* abgleichen. Dies war bei dieser Query allerdings unvorteilhaft, da es sich bei Texten mit diesem Kompo-

situm um ein anderes Thema handelte. Das Auffinden der Texte hatte einen negativen Einfluss auf die Precision, da hauptsächlich nicht relevante Texte gefunden wurden.

Fazit

Der Einfluss morphologischer Werkzeuge auf die Qualität des Information Retrieval wurde in dieser Studie an der Baseline, dem Porter-Stemmer und drei Varianten des *wmtrans*-Lemmatizers getestet. Zwei waren um die Fähigkeit der Kompositazerlegung erweitert. Als Baseline diente eine Version, die lediglich einen Wortabgleich identischer Wörter durchführt.

Die Qualität der Ergebnisse ist abhängig vom Informationsbedürfnis, beziehungsweise von der Komplexität der gewählten Suchbegriffe. Bei komplexen Suchbegriffen, wie zum Beispiel Komposita, sind die beiden *wmtrans*-Lemmatizer, welche um die Eigenschaft der Kompositazerlegung erweitert wurden, in den meisten Fällen überlegen. Die beiden um die Fähigkeit der Wortzerlegung erweiterten wmtrans-Lemmatizer verbesserten den Recall-Wert. Die Precision wurde allerdings verschlechtert.

Literatur

Canoo Engineering AG (2012). Produkt-Homepage: Canoo LanguageTool (ehemals wmtrans) 2012 – URL: http://www.canoo.com/languageexperts/products/toolkit/?l=de – Zugriffsdatum: 29.03.2013.

[CLEF 2003] BRASCHLER, Martin; GONZALO, Julio; KLUCK, Martin (2003)(Hrsg.). Advances in Cross-Language Information Retrieval. Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Rome, Italy, September 19-20, 2002. Revised Papers. Springer Verlag, 2003.

DICK, Melanie (2012). Einfluss morphologischer Werkzeuge auf die Information-Retrieval-Qualität 2012 – Bachelorarbeit. Universität Hildesheim.

PORTER, Martin F. (1997). An algorithm for suffix stripping. In: SPARCK JONED, Karen (Hrsg.); WILLETT, Peter (Hrsg.): Readings in information retrieval. San

Francisco, CA, USA: Morgan Kaufmann Publisher Inc. 1997, S. 313-316. - URL: http://tartarus.org/martin/PorterStemmer/def.txt - Zugriffsdatum: 29.03.2013.