

Seminarprojektarbeit: Extraktion von Zeitleisten aus Wikipedias Succession-Tables

Armin Schmidt

HS Webbasierte Informationsextraktion (Simone Ponzetto), WS 08/09
Seminar für Computerlinguistik, Universität Heidelberg

15. März 2009

1 Einleitung

Wikipedia bietet ein Fülle an semistrukturierter Information und hat sich in jüngster Zeit als wichtige Resource für die sprachverarbeitende Forschung etabliert. So nutzen z. B. Wu and Weld [2007] die in Wikipedia-Infoboxen vorhandene Information und deren Struktur, um Prädikat-Argument-Relationen zu extrahieren. Um ein weiteres Beispiel zu nennen, bauen Ponzetto and Strube [2007] anhand der hierarchischen Kategorisierung von Wikipedia-Artikeln eine Taxonomie auf. Die Herausforderungen, denen sich alle diese Ansätzen stellen müssen, liegen in der gemeinschaftsbasierten Natur der Wikipedia begründet: die Struktur von ganzen Artikeln oder auch nur den in ihnen enthaltenen Elementen ist nur soweit spezifiziert, dass die Kreativität der Nutzer nicht eingeschränkt wird. Strukturelle Vorgaben werden absichtlich nicht enforciert und auch ihnen nicht entsprechende Artikel werden veröffentlicht. Die große Masse an niedrigfrequentierten Artikeln ist aber oft unvollständig und fehlerhaft. Gerade Fehler im MediaWiki-Quellcode, die anhand des im Browser dargestellten HTML-Codes nicht sichtbar sind, werden nur langsam korrigiert. Templates werden, je nach Sichtweise, kreativ eingesetzt bzw. zweckentfremdet. Soll Wikipedia nicht lediglich als Textkorpus genutzt werden, sondern als Quelle strukturierter Information, ist also immer ein nicht zu unterschätzender Aufwand für die Vorverarbeitung einzuplanen.

Für diese Seminarprojektarbeit habe ich ein Programm geschrieben, welches die in vielen Wikipedia-Artikeln enthaltenen sogenannten *Succession-Tables* (deutsch: Nachfolgetabellen) verarbeitet und für weiterführende Methoden der Informationsextraktion vorbereitet. *Succession-Tables* finden sich insbesondere in Wikipedia-Artikeln, die Personen beschreiben und stellen in einer Tabelle die Art und Dauer der von der jeweiligen Person bekleideten Ämter, sowie die Vorgänger/innen dar. Abbildung 1 zeigt ein Beispiel.

S-start House of San Miguel Cadet branch of the House of López Born: 19 January 1958 Died: Living		
Regnal titles		
New title Dowry from father	Queen of India 1972 – 1983 <i>with Michael (1972 – 1975)</i>	Succeeded by Arnold
Preceded by Elizabeth	Lady Supreme of Oceana 1975 – 1983	Vacant Title next held by Jilian
	Empress of Arabia 1975 – 1993	Title merged with Great Khanna of Asia
	Grand Duchess of Europa 1975 – 1999	Succeeded by Felicia
Vacant Title last held by Karl von Igorstein	Chief Sultana of Africa 1993 – 1999	
New title Consolidation of Asia	Great Khanna of Asia 1993 – present	Incumbent <i>Designated heir:</i> <i>Marcus</i>
Preceded by Diane	– TITULAR – Lady of the Isles 2003 – present Reason for succession failure: Declared a republic	

Abbildung 1: Beispiel für einen *Succession-Table*, der eine Großzahl der möglichen Elemente verwendet. Entnommen von: <http://en.wikipedia.org/wiki/Template:S-start>

2 Verarbeitung von Succession-Tables

Bei der Verarbeitung der im MediaWiki-Format vorliegenden *Succession-Tables* sind folgende Umstände und Schwierigkeiten zu berücksichtigen: Für das Anlegen eines *Succession-Tables* innerhalb eines Wikipedia-Artikels stehen zwei *Templates* zur Verfügung. Das als *deprecated* (veraltet) geltende ältere Format ist für sehr einfache Tabellen gedacht, bei denen ein Amt, Amtsperiode, Vorgänger und Nachfolger in einer Zeile dargestellt werden können, und Besonderheiten dieses Amtes weitgehend unvorhanden sind. Ein Beispiel für dessen Syntax ist in Abbildung 2 gegeben. Mit dem neueren Format können mitunter sehr komplexe Abfolgebeziehungen dargestellt werden, was oft bei Erbschaftsfolgen in Herrschaftshäusern notwendig sein kann. Der Quellcode für das anhand des neuen Templates erstellte Beispiel in Abbildung 1 ist in Abbildung 3 gegeben. Beide Formate sind weitestgehend standardisiert und mehr oder weniger vollständig auf den Dokumentationsseiten der Wikipedia beschrieben¹.

Die Herausforderungen eines Programms zur Verarbeitung von *Succession-Tables* lassen sich in den folgenden Punkten zusammenfassen:

1. getrennte Verarbeitung beider *Template*-Formate
2. Entwicklung eines Algorithmus zur korrekten Verarbeitung von mehrzeiligen Zellen und Subtemplates, die in verschiedenen Spalten vorkommen können

¹http://en.wikipedia.org/wiki/Template:Succession_box
<http://en.wikipedia.org/wiki/Template:S-start>

```

{{s-start}}
{{succession box
| before = [[Clement Attlee]]
| title = [[Prime Minister of the United Kingdom]]
| years = 26 October 1951 – 7 April 1955
| after = [[Anthony Eden|Sir Anthony Eden]] }}
{{succession box
| before = [[Emanuel Shinwell]]
| title = [[Minister of Defence (UK)|Minister of Defence]]
| years = 1951 – 1952
| after = [[Harold Alexander, 1st Earl Alexander of Tunis|The Earl Alexander of
Tunis]] }}
{{end}}

```

Abbildung 2: Succession-Table - altes Format

```

{{s-start}}
{{s-hou|House of San Miguel|19 January|1958||Living|House of López|Ricardo III of
Eurasia and I of Africa}}
{{s-reg|}}
{{s-new|reason=Dowry from father}}
{{s-ttl|title=Queen of India|years=1972 1983|regent1=Michael|years1=1972 1975}}
{{s-aft|after=Arnold}}
{{s-bef|rows=3|before=Elizabeth}}
{{s-ttl|title=Lady Supreme of Oceana|years=1975 1983}}
{{s-vac|next=Jilian}}
|-
{{s-ttl|title=Empress of Arabia|years=1975 1993}}
{{s-non|reason=Title merged with<br />Great Khanna of Asia}}
|-
{{s-ttl|title=Grand Duchess of Europa|years=1975 1999}}
{{s-aft|rows=2|after=Felicia}}
{{s-vac|last=Karl von Igorstein}}
{{s-ttl|title=Chief Sultana of Africa|years=1993 1999}}
|-
{{s-new|reason=Consolidation of Asia}}
{{s-ttl|title=Great Khanna of Asia|years=1993 present}}
{{s-inc|rows=2|heir=Marcus}}
|-
{{s-bef|rows=1|before=Diane}}
{{s-tul|title=Lady of the Isles|years=2003 present|reason=Declared a republic}}

```

Abbildung 3: Succession-Table - neues Format

3. Entwicklung von regulären Ausdrücken, welche die gewünschte Information (z. B. Amtsperiode, Vorgänger, Nachfolger) extrahieren und dabei die relativ freien Eingabemöglichkeiten für die einzelnen Felder berücksichtigen

Punkt 3 der vorangegangenen Aufzählung bezieht sich insbesondere auf die Möglichkeiten für einen Autor, den Wert für ein bestimmtes Feld als wikipediainternen Link zu kodieren, letzteres mit oder ohne separaten Ankertext, etc. Bei Datumsfeldern kann ein Wert ebenfalls frei eingegeben werden, d. h. alle der folgenden Formen können vorkommen:

- 2000 - 2003
- 2000 - Spring 2003
- 1 September 2000 - 1 March 2003
- 2000
- 1. 9. 2000 - 1. 3. 2003
- 2000 - present
- [[1 September 2000|2000]] - [[1 March 2003|2003]]
- evtl. andere

Die zuletzt genannte Schwierigkeit wird in der jetzigen Version der Implementierung jedoch größtenteils umgangen, indem lediglich Jahresangaben extrahiert werden.

3 WPTimelineExtractor

WPTimelineExtractor ist eine Java-Implementierung zur Verarbeitung von Wikipedias Succession-Tables. Das Programm führt folgende Schritte durch:

1. Andocken an einen Wikipedia-MySQL-Dump. Die Extraktion einzelner Artikel im MediaWiki-Format erfolgt mit Hilfe der Bibliothek JWPL (Zesch et al., 2008).
2. Extraktion von *Succession-Tables* und der darin enthaltenen Informationsfelder.
3. Speichern einer einzelnen Abfolgeeinheit in einem *Succession*-Objekt. Die Java-Klasse *Succession* ist eine reine Datenklasse und enthält Felder zum Speichern von Artikelüberschrift (eindeutig einem bestimmten Artikel zuzuordnen, und damit gleichzeitig eine Art Identifikationseinheit für Artikel), Verweis und Ankertext für Amtstitel, Dauer des besetzten Amtes, Vorgänger sowie Nachfolger.
4. Überführung aller *Succession*-Objekte in eine Abbildung von Amtstiteln auf jeweils eine sortierte Liste von diesen Titeln entsprechenden *Succession*-Objekten.

5. Persistentes Abspeichern jener Abbildung in eine Datei sowohl im Binär- als auch im Textformat.

WPTimelineExtractor wird unter der GNU Lesser Public License veröffentlicht und kann unter <http://armin.diotavelli.net/WPTimelineExtractor> frei heruntergeladen werden. Das Programm wird von der Kommandozeile gestartet und verlangt einen Parameter, der die maximal zu extrahierende Anzahl von Succession-Objekten beschreibt. Ist dieser Parameter gleich -1, so wird der gesamte Wikipedia-Dump traversiert. Der folgende Befehl beispielsweise extrahiert die ersten 100 Successions und speichert diese als `timelines.dump` und `timelines.txt` im aktuellen Verzeichnis (der Pfad zu JWPL muss dabei angegeben werden):

```
> java -classpath wptimelinesextactor.jar:\
jwpl_0.45beta.jar:.de.uni.heidelberg.cl.schmidar.wptimelines.TimelineExtractor 100
```

4 Abbilden von Amtstiteln

Eine Zeitleiste ist eine Abbildung eines Amtstitels auf eine geordnete Liste von Successions. Entsprechen die den Amtstitel beinhaltenden Felder zweier Succession-Objekte einander, so gehören sie derselben Zeitleiste an. Dieses grundlegend einfache Prinzip führt in der Praxis jedoch zu Herausforderungen, von denen einige hier kurz umrissen werden sollen:

1. Ist ein Amtstitel nicht mit einem diesen Titel beschreibenden Artikel verlinkt, so kann kaum sichergestellt werden, dass die Titelbezeichnung korrekt ist. Sind in zwei verschiedenen Artikeln die hypothetischen Amtstitel *King of Italy* und *Ruler of Italy* als einfache, unverlinkte Strings gegeben, so gibt es keinen direkten Weg, beide Bezeichnungen auf dasselbe Amt abzubilden. Zur Lösung dieses Problem könnte ein Ansatz dienen, der *Ruler* und *King* als quasi-synonym erkennt, indem z. B. deren Relation in einer Taxonomie wie WordNet festgestellt wird.
2. Bei verlinkten Titelangaben mit unterschiedlichen Ankertexten können die Titel der Zielseiten zur Disambiguierung dienen (d. h. *King of Italy* und *Ruler of Italy* würden auf dieselbe Seite verweisen). Selbst Links auf denselben Artikel können jedoch aufgrund des *Redirection*-Mechanismus von Wikipedia unterschiedliche Linktexte besitzen. In diesem Fall könnte der Link jedoch einfach bis zur Zielseite verfolgt werden, so dass jener Artikel als Identifizierer gilt. Negative Auswirkungen auf die Performanz sind hierbei zu erwarten.
3. Gibt es keinen Artikel für ein bestimmtes Amt, so wird oft auf einen Artikel verlinkt, dessen Thema mit dem jeweiligen Amt in Verbindung steht. Der Artikel über Angela Merkel beispielsweise führt unter anderen die drei zweifelsfrei unterschiedlichen Ämter *Secretary General of the Christian Democratic Union of Germany*, *Chairwoman of the Christian Democratic Union of Germany* sowie *Chairwoman of the CDU/CSU parliamentary group* auf. Jedes der drei Ämter verweist jedoch auf den Artikel *Christian Democratic Union (Germany)*. Der Ankertext kann zwar zur Disambiguierung verwendet werden, stellt jedoch keinen eindeutigen Identifizierer dar und konfrontiert uns mit dem in Punkt 1 beschriebenen Problem. Eine Methode, die zumindest erkennt, ob ein Link richtig sein kann oder nicht, könnte von der Kategorisierung von Wikipedia-Artikeln zehren. Demnach sollte z.B. *Chairwoman of the Christian Democratic Union of Germany* nicht auf einen Artikel verweisen, der unter *Political Parties in Germany* kategorisiert ist.

Literatur

- S.P. Ponzetto and M. Strube. Deriving a large scale taxonomy from wikipedia. In *Proceedings of the national conference on artificial intelligence*, volume 22, page 1440. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.
- F. Wu and D.S. Weld. Autonomously semantifying wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50. ACM New York, NY, USA, 2007.
- Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, 2008.