

MFG-STIFTUNG BADEN WÜRTTEMBERG
KARL-STEINBUCH-STIPENDIUM

HeiNER - the Heidelberg Named Entity Resource

ABSCHLUSSBERICHT FÜR DAS KARL-STEINBUCH-STIPENDIUM
2008

Autoren:

Johannes KNOPP
Carina SILBERER
Wolodja WENTLAND

Betreuung:

Prof. Dr. Anette FRANK

Email:

HEINER@CL.UNI-HEIDELBERG.DE

Datum: 17. Dezember 2008

Inhaltsverzeichnis

1	Einleitung	3
1.1	Motivation	3
1.2	HeiNER	3
1.3	Wikipedia	4
1.3.1	Aufbau von Wikipedia	5
2	Projektverlauf	7
2.1	Vorarbeiten	8
2.1.1	Zugriff auf Wikipedia Daten	8
2.1.2	Datenextraktion	9
2.2	Projektarbeit	11
2.2.1	Wikipedia API	11
2.2.2	Datenextraktion	12
2.2.3	Evaluierung der Daten	15
2.2.4	Ausarbeitung der Publikation	18
2.2.5	Erstellen des Posters	18
2.2.6	LREC 2008	19

1 Einleitung

Das Projekt hatte zum Ziel, automatisch mithilfe von Wikipedia eine große multilinguale Datenbank mit dem Titel *HeiNER – the Heidelberg Named Entity Resource* zu erstellen, die Named Entities (Eigennamen) mit ihren sprachlichen Kontexten sowie Transliterationen und Übersetzungen in einer Vielzahl von Sprachen enthält.

1.1 Motivation

Im Regelfall können Named Entities nicht in Wörterbüchern nachgeschlagen werden, automatische Sprachverarbeitungssysteme betrachten sie deshalb als unbekannte Wörter. Es sind jedoch Named Entities, die besonders viel über den Inhalt eines Textes aussagen, da vorkommende Personen, Organisation oder Orte das Thema des Textes aufschlüsseln können.

Für viele Anwendungen der Computerlinguistik sind daher Verfahren notwendig, die Named Entities erkennen und ihrer korrekten semantischen Klasse zuordnen. Moderne Methoden zur Erkennung und Klassifikation beruhen auf statistischen Lernverfahren. Diese benötigen große Mengen von Trainingsdaten, die manuell mit erheblichem Aufwand für jede Sprache einzeln aufbereitet werden müssen und aus diesem Grund nur begrenzt zur Verfügung stehen. Deshalb erreichen viele Lernsysteme ihre maximale Leistung noch nicht. Signifikante Verbesserungen verspricht man sich durch Verfahren, die automatisch große Mengen von Trainingsdaten generieren können.

Unser System verfolgt genau diesen Ansatz, indem es zuerst Named Entities aus der Online-Enzyklopädie Wikipedia für eine Sprache extrahiert, diesen dann Transliterationen (Übertragungen in andere Schriftsysteme) bzw. Übersetzungen zuordnet und als beispielhafte sprachliche Umgebung für jede Named Entity Sätze speichert, in denen diese vorkommt. Dazu nutzen wir die Eigenschaft von Wikipediaartikeln, dass sie über Sprachen hinweg miteinander verknüpft sind. Wir bilden die transitive Hülle über die Menge der Sprachlinks und erhalten ein vielsprachiges Wörterbuch. Die so gewonnenen Daten können zum Training von Lernsystemen oder auch zur Erweiterung bestehender Wissensbasen dienen.

1.2 HeiNER

Das Ziel des Projektes war die Erstellung und Bereitstellung folgender Komponenten, die ausschließlich aus der Online-Enzyklopädie Wikipedia gewonnen wurden:

- ein Wörterbuch mit den Übersetzungen von Named Entities der englischen Sprache

in alle Sprachen, die in Wikipedia als Sprachversion zur Verfügung stehen (Details s. 2.2.2)

- für jede Zielsprache eine Datenmenge von Kontexten der Named Entities (Details s. 2.1.2)

In Unterabschnitt 2.2.3 sind die qualitativen und quantitativen Resultate dargestellt.

1.3 Wikipedia

Wikipedia ist eine auf MediaWiki¹ basierende Online-Enzyklopädie, deren Inhalte von Benutzern der Enzyklopädie in gemeinschaftlicher Arbeit erstellt werden. Die Erstellung eines Artikels erfolgt ausschließlich durch aufeinander aufbauende Änderungen unterschiedlicher Benutzer, die angehalten werden, sich an die im *Manual of Style*² festgelegten Richtlinien zu halten.

Wikipediadaten erfreuen sich seit einiger Zeit eines hohen Interesses seitens der computerlinguistischen Forschungsgemeinschaft und wurden bereits erfolgreich in einer Reihe von Gebieten eingesetzt, darunter:

- Maschinelle Übersetzung (Alegria et al., 2006)
- Named Entity Transliteration (Sproat et al., 2006)
- Word Sense Disambiguation (Mihalcea, 2007)
- Konstruktion paralleler Korpora (Adafre and de Rijke, 2006)
- Konstruktion von Ontologien (Ponzetto and Strube, 2007)

Aus computerlinguistischer Sicht besteht die Attraktivität von Wikipedia insbesondere in der hohen Anzahl von Artikeln zu Named Entities, was Wikipedia maßgeblich von traditionell eingesetzten Ressourcen wie WordNet (Fellbaum, 1998) unterscheidet.

Wikipedia ist in den letzten Jahren in beeindruckender Weise gewachsen und umfasste im September 2007 etwa 9,25 Millionen Artikel in 253 Sprachen mit insgesamt über 1,74 Milliarden Wörtern.

Aktuelle Statistiken zeigen, dass Wikipedia weiterhin großes Wachstum verzeichnen kann und somit eine ideale Grundlage zur Erstellung von umfangreichen, aktuellen und nachhaltigen Ressourcen wie *HeiNER* bietet.

¹<http://www.mediawiki.org>

²http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style

1.3.1 Aufbau von Wikipedia

Wikipedia ist ein großes Hypertext Dokument welches aus einzelnen Artikeln, oder Seiten, besteht die untereinander mit Links verbunden sind. Die Mehrheit dieser Links dient nicht nur der Navigation innerhalb von Wikipedia, sondern vor allem dem Verweis auf Artikel, welche die verlinkte Entität oder das Konzept diskutieren. Eine für unser Projekt wichtige Eigenschaft von Artikeln in Wikipedia ist es, dass diese eindeutige Entität diskutieren, bzw klar ausdrücken, welcher Wortsinn gemeint ist.

Markup

Wikipedia ist eine auf der freien Wiki Software MediaWiki basierende Enzyklopädie und bedient sich der für MediaWiki definierten Markup-Sprache. MediaWiki-Markup wurde leider nicht formal spezifiziert und es existiert kein offizieller Parser der zur Verarbeitung eingesetzt werden könnte. Die Markup Sprache wurde vielmehr sukzessive erweitert um dem Wunsch nach neuen Funktionen des MediaWikis nachzukommen. Unser Projekt versucht möglichst alle Funktionalitäten abzudecken, die in der offiziellen Wikipedia Dokumentation beschrieben sind, doch in vielen Fällen mussten wir den MediaWiki Code analysieren um zu verstehen, wie ein bestimmter Fall wirklich behandelt wird.

Links

Links innerhalb von Wikipedia stellen eine Referenz auf einen anderen Artikel innerhalb eines Textes dar. Links haben in Wikipedia die generelle Form `[[Titel]]` und können zusätzlich noch den im Artikel angezeigten Ankertext angeben und haben dann die Form `[[Titel | Ankertext]]`. Neben diesen einfachen Linktypen finden sich noch eine Vielzahl an spezielleren Links, die der Einbettung von multimedialen Inhalten, der Zuweisung von Kategorien oder der multilingualen Verlinkung dienen.

Links zwischen Sprachversionen

Links zwischen Sprachversionen verlinken einen Artikel über eine Entität oder ein Konzept in einer Sprache mit dem entsprechenden Artikel in einer anderen Sprache. Links zwischen einzelnen Sprachen haben die Form `[[Sprachkennung:Titel]]`.

Seiten

Seiten in Wikipedia können theoretisch in vier Klassen aufgeteilt werden: *Konzept-Seiten*, *Named-Entity-Seiten* sowie *Kategorie-Seiten* bilden die Mehrheit der in Wikipedia vorhandenen Seiten. Die Unterscheidung zwischen Konzepten und Named Entities findet sich nicht in der Struktur von Wikipedia wieder und stellt somit eine der Hauptaufgaben unseres Projektes dar. Die vierte Klasse bilden die *Meta Seiten*, die der Einbindung von unterschiedlichen Medien, administrativen Aufgaben oder der Disambiguierung von ambigen Bezeichnern dienen. Jede Seite in Wikipedia hat einen eindeutigen Titel, wie zum Beispiel `Python(programming language)` und einen zugehörigen Text.

Redirect-Seiten

Redirect-Seiten dienen der Normalisierung von unterschiedlichen Oberflächenformen, Be-

zeichnungen oder morphologischen Varianten einer bestimmten nicht ambigen Named Entity oder eines Konzepts zu einem eindeutigen Bezeichner. Redirect-Seiten werden zum Beispiel benutzt, um Titel wie *USA*, *United States of America* und *Yankee Land* auf den Artikel *United States* weiterzuleiten. Diese Redirect-Seiten ermöglichen die Erkennung von *USA*, *United States of America* und *Yankee Land* als unterschiedliche Oberflächenformen der Named Entity *United States*.

Der Titel einer Redirect-Seite ist die zu normalisierende Zeichenkette und der Text gibt das Ziel der Weiterleitung an. Der Text einer Redirect-Seite hat immer die nachfolgend angegebene Form, was die Erkennung dieser Seiten erheblich vereinfacht.

```
#REDIRECT [[Redirect goal]] {{Redirect reason}}
```

Disambiguierungsseiten

Disambiguierungsseiten dienen der Auflistung aller eindeutigen Named Entities die durch einen ambigen Eigennamen bezeichnet werden können. Die Disambiguierungsseite für *Python* enthält zum Beispiel Links auf Artikel, die *Pythonidae*, die englische Komikergruppe *Monty Python* oder die Programmiersprache *Python (programming language)* diskutieren.

2 Projektverlauf

Dauer	Tätigkeit
Fünf Wochen	Anpassungen zum Umgang mit zahlreichen Sprachen: Wikipedia-API (Markup-Übersetzung) Datenextraktionsprogramme
Vier Wochen	Modifikation und Erweiterung der schon bestehenden Programme (s. 2.2.2)
Drei Wochen	Ausführen der Programme, Extraktion der Daten für <i>HeiNER</i>
Drei Wochen	Evaluierung der Daten, Ausarbeiten der Publikation
Drei Wochen	Erstellen des Posters für die Konferenz
Eine Woche	Besuch der Language Resources and Evaluation Conference (LREC) 2008
Zwei Wochen	Bugfixing (Vorher ausgelassene Sprachen hinzugefügt)
Drei Wochen	Neugenerierung der Daten und Grafiken
Zwei Wochen	Aufbau der Webseite (http://heiner.cl.uni-heidelberg.de), freie Veröffentlichung von <i>HeiNER</i>

Tabelle 2.1: Verlauf des Projekts

Da die Idee für *HeiNER* aus einem Softwareprojekt im Sommersemester 2007 entstanden ist, konnten wir zum Projektbeginn auf zuvor von uns implementierte Software zurückgreifen. Diese umfasst eine Benutzerschnittstelle (API) zur englischen Wikipedia, die Erkennung der dort enthaltenen Named Entities und das Extrahieren der zugehörigen Kontexte. Dies entspricht der einsprachigen Variante von *HeiNER*. Tabelle 2.1¹ gibt einen Überblick über den Verlauf des geförderten Projektes. Detailliertere Angaben der Inhalte folgen im Abschnitt 2.2.

¹Die Angaben sind als generelle Zeit- und Aufwandseinordnung zu betrachten. Viele Arbeitsschritte haben sich überschritten oder mussten auf die Ergebnisse anderer Schritte warten. Die Geschwindigkeit von Programmausführungen variierte, je nachdem mit wievielen weiteren Benutzern wir den Server des Seminars für Computerlinguistik teilen mussten, der uns zur Verfügung stand.

2.1 Vorarbeiten

2.1.1 Zugriff auf Wikipedia Daten

Die im Rahmen des Softwareprojektes entwickelte API lässt sich grob in folgende Module aufteilen:

- Zugriff auf Wikipediadaten
- MediaWiki Markup Verarbeitung
- Objektmodell für Wikipedia Artikel

Wikipediadaten stehen in Form eines in XML serialisierten Datenbankdumps zur Verfügung. Die Problematik bei der Verarbeitung eines solchen Dumps besteht weniger in seiner komplexen Struktur, als vielmehr in der Größe der zu verarbeitenden Dateien².

Der folgende exemplarische Auszug aus dem XML-Dump der englischen Wikipedia zeigt die Elemente innerhalb des Dumps, die für unser Projekt relevant waren.

```
<mediawiki ... >
  ...
  <page>
  <title>Pi</title>
  ...
  <revision>
    ...
    <text xml:space="preserve">Pi is a mathematical constant that
      represents the ratio of any circle 's circumference to its diameter in
      ... </text>
  </revision>
</page>
<page>
  ...
```

Da wir nur an den Titeln und dem Inhalt von Artikeln, aber nicht an Metadaten, wie zum Beispiel deren Erstellungsgeschichte, interessiert waren, mussten wir nur eine geringe Teilmenge der im XML-Dump gespeicherten Information bearbeiten.

Zu diesem Zweck wurde eine Klasse (`WPDumpExtractor`) geschrieben, welche die XML-Dumps zeilenweise bearbeitet, relevante Passagen mit Hilfe von einfachen regulären Ausdrücken erkennt und die extrahierten Daten in Form von Artikelobjekten zur Verfügung stellt.

Die notwendige Differenzierung zwischen einzelnen Seitentypen (Bilder-, Medien-, . . . , Disambiguierungsseiten) wurde anhand typischer Zeichenketten in den Titeln der betrachteten Seiten vorgenommen. So konnten Seiten für Bilder zum Beispiel an dem Präfix `Image:` erkannt werden.

²Der XML-Dump der englischen Wikipedia hatte Ende 2007 eine Größe von etwa 13GB

```

<context id='198'>
  <surfaceForm>WWF</surfaceForm>
  <leftContext>According to the</leftContext>
  <rightContext>the territory of Belgium belongs to the ecoregion
    of Atlantic mixed forests.</rightContext>
  <ne>World Wide Fund for Nature</ne>
</context>
...
<context id='9568'>
  <surfaceForm>Wiener Prater</surfaceForm>
  <leftContext>There are also two miniature railways:the
    Liliputbahn in the</leftContext>
  <rightContext>and the Donauparkbahn in the Donaupark.</rightCtxt>
  <ne>Prater</ne>
</context>

```

Abbildung 2.1: Beispiel für einen *context* im *context dataset*

2.1.2 Datenextraktion

Heuristik zur Erkennung von Named Entities im Englischen

Named Entities haben im Englischen die Eigenschaft großgeschrieben zu werden, was die Aufgabe der Named Entity Recognition (NER) vereinfacht. Zur Erkennung von Named Entities in der englischen Wikipedia haben wir eine von (Bunescu and Pasca, 2006) vorgestellte Heuristik nachimplementiert.

Kontextextrahierung

Die Implementierung des in diesem Abschnitt vorgestellten *dataset_{ambiguous}* und des dafür notwendigen *disDict* orientiert sich an der Arbeit von (Bunescu and Pasca, 2006).

Ein Kontext(eintrag) einer Named Entity *ne* besteht neben dem Paragraphen, in dem *ne* auftritt, aus der Zeichenkette, mit der im Paragraphen auf *ne* referiert wird, sowie aus *ne* selbst. Die Menge aller Kontexte einer gegebenen Sprache bilden ein (*context*) *dataset*.

Die Erstellung eines *datasets* erfolgt durch den sogenannten *dataset_builder*. Dieser iteriert über alle Artikel einer Wikipedia-Sprachversion. Für jede Named Entity, die **verlinkt** in einem Artikel vorkommt, wird der die Named Entity umgebende Paragraph als deren Kontext extrahiert. Eine Named Entity *ne* tritt verlinkt in einem Artikel auf, wenn das Ziel eines Links *l* der Artikel mit dem Titel *ne* ist.

Bevor nun die genaue Bedingung genannt wird, die auf Links für die Extrahierung ihres Kontextes zutreffen muss, wird im Folgenden zunächst eine dafür grundlegende Komponente erläutert.

Disambiguierungswörterbuch

Für die Extrahierung der Kontexte bildet das Disambiguierungswörterbuch (*disDict*)

```

<dictSet>
  <pageTitle>WWF</pageTitle>
  <ne>World Wide Fund for Nature</ne>
  <ne>World Wrestling Entertainment</ne>
  <ne>World Water Forum</ne>
  <ne>Wesley Willis Fiasco</ne>
  <ne>Windows Workflow Foundation</ne>
</dictSet>
...
<dictSet>
  <pageTitle>Stuttgart</pageTitle>
  <ne>Stuttgart</ne>
  <ne>Stuttgart, Arkansas</ne>
  ...
</dictSet>

```

Abbildung 2.2: Beispiel für zwei Einträge im *disDict*

eine Grundlage.

Das *disDict* bildet die Relation zwischen Eigennamen und den Named Entities, auf die sie verweisen können, ab. Abb. 2.2 ist zum Beispiel zu entnehmen, dass der Eigenname WWF (der Schlüssel des Eintrages) auf die Named Entities **World Wide Fund for Nature** oder **World Water Forum** (Werte des Eintrages) verweisen kann. Die Erstellung des Disambiguierungswörterbuches umfasst die folgenden Schritte:

1. Für jede Named Entity *ne*:
Füge einen Eintrag in *disDict* ein, mit *ne* als Schlüssel und als Wert. Dies spiegelt die Tatsache wider, dass jede Named Entity auf sich selbst verweisen kann.
2. Für jede Disambiguierungsseite *d*, die mindestens einen Link auf eine Named Entity *ne* beinhaltet:
Füge dem Eintrag mit dem Schlüssel *d* alle Named Entities, auf die *d* verlinkt, als Werte hinzu.
3. Für jede Redirect-Seite *r*, die auf eine Named Entity *ne* verlinkt:
Füge dem Eintrag mit dem Schlüssel *r* den Wert *ne* hinzu.

Versionen von *datasets*

Je nach Verwendungszweck der Kontexte können zwei Versionen unterschieden werden: ein allgemeines *dataset_complete* und ein ambiges *dataset_ambiguous*. Während das *dataset_complete* die Menge aller Kontexte von auf Named Entities verweisenden Links ist, beinhaltet das *dataset_ambiguous* nur solche Kontexte, die per se mehrdeutig sind.

Die zusätzliche Einführung des *dataset_complete* wurde erst während der Förderungszeit getroffen, zuvor war lediglich die Implementierung des *dataset_ambiguous* vorhanden. Die Erstellung des *dataset_complete* wird deshalb in Unterabschnitt 2.2.2 erläutert.

Die Mehrdeutigkeit des *dataset_ambiguous* beruht auf den Ankertexten der Links: Kann ein Ankertext (Eigenname) generell auf mehr als eine bestimmte Named Entity verweisen, so ist dieser mehrdeutig und somit auch der gegebene Kontext. Da jedoch das Ziel des Links, d.h. die Named Entity, im Link angegeben ist, sind die Kontexteinträge im *dataset* durch das explizite Angeben des Ziels in Form von `<ne>Ziel</ne>` eindeutig.

Für die Identifizierung eines mehrdeutigen Links wird das *disDict* verwendet: mehrdeutige Eigennamen werden durch Einträge repräsentiert, bei denen die Menge der Werte, d.h. die Menge der Named Entities, auf die der Eigenname referieren kann, mindestens aus zwei Elementen, i.e. Named Entities, besteht.

Die Bedingung für die Extrahierung eines Kontextes für das *dataset_ambiguous* wird somit folgendermaßen formuliert:

- Der Ankertext von *l* ist der Schlüssel eines Eintrages im *disDict*. Die zugehörige Wertemenge besteht aus mindestens zwei Named Entities, wobei eine von ihnen das Ziel von *l* ist.

Ein Beispiel für einen mehrdeutigen Kontext bildet der erste Eintrag in Abb. 2.1.

2.2 Projektarbeit

Zum Aufbau aller Komponenten von *HeiNER* mussten die bereits bestehenden Programme für die multilinguale Verarbeitung von Wikipedia erweitert werden. Die nötigen Anpassungen werden in den folgenden Abschnitten vorgestellt.

2.2.1 Wikipedia API

Zu Beginn des Projekts nahmen wir eine umfassende Analyse der bisher von uns erzeugten monolingualen Daten vor, die eine Reihe von Problemen in den folgenden Bereichen offenbarte:

- Erkennung und Behandlung von HTML-Entities
- Verarbeitung von komplexen Linktypen

Wikipediaartikel enthalten HTML-Entitäten wie `ä`; die wir bisher nicht separat erkannt und verarbeitet hatten, so dass wir einen zusätzlichen Verarbeitungsschritt zur Ersetzung dieser Entitäten durch das entsprechende Zeichen einfügen mussten.

Die von uns implementierte API konnte nur einfache Linktypen wie `[[Ziel]]` und `[[Ziel | Ankertext]]` explizit verarbeiten, was zu einer unvollständigen Entfernung des Link Markups oder fehlenden Textfragmenten geführt hat. Dieses Problem lösten wir durch die Implementierung einer umfassenden Klassenbibliothek für alle in Wikipedia verwendeten Linktypen.

Die Erweiterungen der API, die notwendig waren, um mit unterschiedlichen Sprachversionen von Wikipedia umgehen zu können waren:

- Erkennung und Verarbeitung von Sprachlinks
- Übersetzung des Markups
- Unterstützung von Namespaces

Die erste notwendige Erweiterung der API war die Implementierung einer korrekten Erkennung und Verarbeitung von Sprachlinks. Die erste Version behandelte nur Sprachlinks der Form `[[cc:Ziel]]` in welcher `cc` für den in Wikipedia vergebenen Sprachcode steht. Dies erfasst nur eine kleine Teilmenge der verwendeten ISO-639 Sprachcodes, so dass wir eine umfassende Liste der verwendeten Sprachcodes erstellt haben und Sprachlinks somit genauer erkennen können.

Die Notwendigkeit auch in anderen Sprachen zwischen unterschiedlichen Seitentypen zu unterscheiden bereitete uns große Probleme, da die zur Kennzeichnung dieser Seitentypen verwendeten Präfixe oder Suffixe leider in den verschiedenen Sprachen in teilweise übersetzter Form verwendet werden. Das in der englischen Version zur Kennzeichnung von Bilderseiten verwendete Präfix `Image:` wird zum Beispiel in der deutschen Version als `Bild:` übersetzt.

Während der Erstellung einer umfassenden Liste dieser Zeichenketten wurde deutlich, dass diese Art der Markup-Verarbeitung einen extremen manuellen Aufwand bei der Portierung und Pflege der API bedeuten würde. Eine eingehende Analyse des MediaWiki Quellcodes zeigte uns eine elegantere Lösung dieses Problems auf. MediaWiki definiert nämlich eine Reihe von Namespaces in welche jeder Artikel eingeordnet wird, so dass eine Erkennung des Seitentypes relativ einfach möglich ist. Insbesondere Templates werden allerdings in allen Sprachen unabhängig definiert und müssen somit leider komplett übersetzt werden.

Diese Änderungen machten eine komplette Neugestaltung der markupverarbeitenden Klassen notwendig, da diese in Abhängigkeit von der Sprache eines Artikels eine unterschiedliche Teilmenge des insgesamt bekannten Markups bearbeiten mussten.

2.2.2 Datenextraktion

(a) Monolingual

Zu Beginn des Projekts bestanden die bereits beschriebenen Programme zur Extraktion der Named Entities in der englischen Sprache und zur Erstellung von *disDicts* (*disDict_builder*) und *datasets* (*dataset_builder*).

Während der Förderungszeit durch die MFG-Stiftung wurden der (*dataset_builder*) um die Unterscheidung von einem mehrdeutigen (*dataset_ambiguous*) und einem allgemeinen (*dataset_complete*) Dataset erweitert sowie Laufzeitverbesserungen durchgeführt. Außerdem wurde dem *disDict_builder* die Erstellung eines *Redirect-Wörterbuches* hinzugefügt, was eine zahlenmäßige Steigerung bezüglich der Kontexteinträge des *dataset_complete* zur Folge hatte:

```

<redirDictSet>
  <pageTitle>United States of America</pageTitle>
  <ne>United States</ne>
</redirDictSet>
...
<redirDictSet>
  <pageTitle>US</pageTitle>
  <ne>United States</ne>
</redirDictSet>

```

Abbildung 2.3: Beispiel für zwei Einträge im *redirDisDict*

Erstellung eines *Redirect-Wörterbuches*

Das *Redirect-Wörterbuch* enthält explizit Abbildungen von Redirect-Seitentiteln auf ihre entsprechenden Named Entity-Artikeltitel (Beispiel s. Abb. 2.3).

Grund für dessen Einführung ist, dass von den in Abschnitt 1.3.1 gegebenen Linkformen noch die folgende Form zu unterscheiden ist: `[[Redirectlink]]`. Bei dieser Form verlinkt der Link auf eine Redirect-Seite, was zur Folge hat, dass man beim Klicken des Links auf die Seite, auf die die Redirect-Seite verlinkt, weitergeleitet wird. Der Titel dieser Seite ist offensichtlich nicht im Link enthalten, somit gibt der Link selbst auch keinen Aufschluss darüber, ob die entsprechende Redirect-Seite auf eine Named Entity verweist.

Das *Redirect-Wörterbuch* wird in das *disDict* integriert. Dies wird durch folgende Modifikation der 3. Regel zur *disDict*-Erstellung (cf. 2.1.2) bewerkstelligt:

3. Für jede Redirect-Seite r , die auf eine Named Entity ne verlinkt:
 - falls in *disDict* ein Eintrag mit r als Schlüssel vorhanden:
Füge dem Eintrag den Wert ne hinzu.
 - sonst:
Füge einen Eintrag ins *redirectDict* ein, mit r als Schlüssel und ne als dessen Wert.

Erstellung des *dataset_complete*

Bei der Erstellung eines *dataset_complete* müssen zur Extrahierung eines Kontext(eintrages) nun eine der beiden Bedingungen zutreffen:

- l verlinkt direkt auf eine der zuvor extrahierten Named Entities, das Ziel von l ist also eine Named Entity, wobei der Ankertext (sofern vorhanden) beliebig sein kann.
- l ist ein Redirect-Link und als Schlüssel im *Redirect-Wörterbuch* enthalten. In diesem Fall gibt das *Redirect-Wörterbuch* Aufschluss über das Ziel von l , d.h. die NE. Die *surfaceForm* ist l .

```

<transDict>
  <namedEntity id='2134'>
    <an>Organizazi3n d'as Nazions Unitas</an>
    <bs>Ujedinjeni narodi</bs>
    <ga>Nisiin Aontaithe</ga>
    <gl>ONU</gl>
    <hu>Egyes3lt Nemzetek Szervezete</hu>
    <lb>Vereent Natiouunen</lb>
    <nds>Vereente Natschonen</nds>
    <tr>Birleŝmiŝ Milletler</tr>
    <en>United Nations</en>
    ...
  </namedEntity>
  ...
</transDict>

```

Abbildung 2.4: Beispiel eines Eintrages im *transDict*

(b) Multilingual

Das Ziel des Projektes war es, eine multilinguale Datenbank zu erstellen, bestehend aus den Kontexten (*datasets*) von Named Entities in verschiedenen Sprachen sowie einem Wörterbuch mit den Übersetzungen und Transliterationen der Named Entities in eine Vielzahl von Sprachen (Beispiel s. Abb. 2.4).

Erstellung des Wörterbuches *transDict*

Um Übersetzungen und Transliterationen von einer Quellsprache in verschiedene Zielsprachen zu erhalten, werden die Sprachlinks von Wikipedia ausgenutzt: Für jede Named Entity, die in der englischen Wikipedia (Quelle) erkannt wurde, werden die zugehörigen Sprachlinks in die anderen Sprachversionen (Ziel) extrahiert.

Die Methode hängt somit stark von der Dichte der Sprachlinkstruktur von Wikipedia ab. Jedoch sind die Sprachlinks eines bestimmten Artikels häufig nicht vollständig, d.h. es bestehen sogenannte *linkage holes* - fehlende Verlinkungen zwischen sich entsprechenden Artikeln in unterschiedlichen Sprachen.

In Abb. 2.5 ist ein Beispiel dafür gegeben. In der Abbildung sind drei Listen von Sprachlinks zu dem Artikel 'Le Crestet' dargestellt: Der englische Artikel verlinkt auf den zugehörigen französischen Artikel, dieser verlinkt u.a. auf den entsprechenden italienischen Artikel. Zwischen dem englisch und italienischen Artikel existiert allerdings keine Verlinkung, es liegt ein *linkage hole* vor.

Um dennoch eine möglichst hohe Abdeckung zu erhalten, wird die transitive Hülle der Sprachlinks einer Named Entity berechnet. Der mittlere Kasten in Abb. 2.6 illustriert dies: Sei eine Named Entity *en* in der englischen Sprache gegeben. Die Sprachlinks des korrespondierenden Artikels liefern sofort die Übersetzung in die deutsche, französische und russische Sprache. Durch das Extrahieren der Sprachlinks des entsprechenden deutschen Artikels kann außerdem die Übersetzung (Transliteration) ins Chinesische gewon-

languages	autre langue
■ Français	■ Français
Autres langues	■ Nederlands
■ Italiano	■ Polski
■ Nederlands	■ Српски / Srpski
■ Polski	■ Volapük
■ Српски / Srpski	
■ Volapük	

Abbildung 2.5: Beispiel eines *linkage holes*

nen werden. Die Untersuchung des zugehörigen chinesischen Artikels liefert eine weitere Übersetzung in die Sprache 'xx'.

Des weiteren werden alle Wikipedia-Sprachversionen der betrachteten Zielsprachen auf Links zu englischen Named-Entity-Artikeln untersucht und ggf. weitere Übersetzungen hinzugefügt.

Die Menge aller Übersetzungen der englischen Named Entities in alle in Wikipedia vorhandenen Sprachen werden im *transDict* gespeichert.

Erstellung der *context datasets* für jede Zielsprache

Für jede betrachtete Zielsprache kann aus dem *transDict* eine Liste von Named Entities gelesen werden. Die Extrahierung der Kontexte ist nun wieder eine monolinguale Aufgabe, d.h. für jede Zielsprache werden *datasets* erstellt, wie es in Abschnitt 2.1.2 erläutert.

2.2.3 Evaluierung der Daten

Die Qualität einer Ressource wie *HeiNER* ist von großer Bedeutung dafür, ob sie für wissenschaftliche Zwecke verwendet werden kann. Die Güte der Daten hängt hauptsächlich von zwei Teilen ab:

1. Die Korrektheit der erkannten Named Entities
2. Die Anzahl der Named Entities, deren Übersetzungen gefunden werden

Named Entities

Um die Güte der erkannten Named Entities zu evaluieren, wurden aus den Daten von *HeiNER* zwei Evaluierungsmengen mit jeweils 2000 Named Entities zufällig ausgewählt. Die erste Menge wurde von zwei, die zweite von drei Annotierern bewertet. Als Annotierrichtlinie galten die Anforderungen des CoNLL-2003 Wettbewerbs für Named Entity Erkennung Sang and Meulder (2003). Um einzuschätzen, wie schwierig die Entscheidung

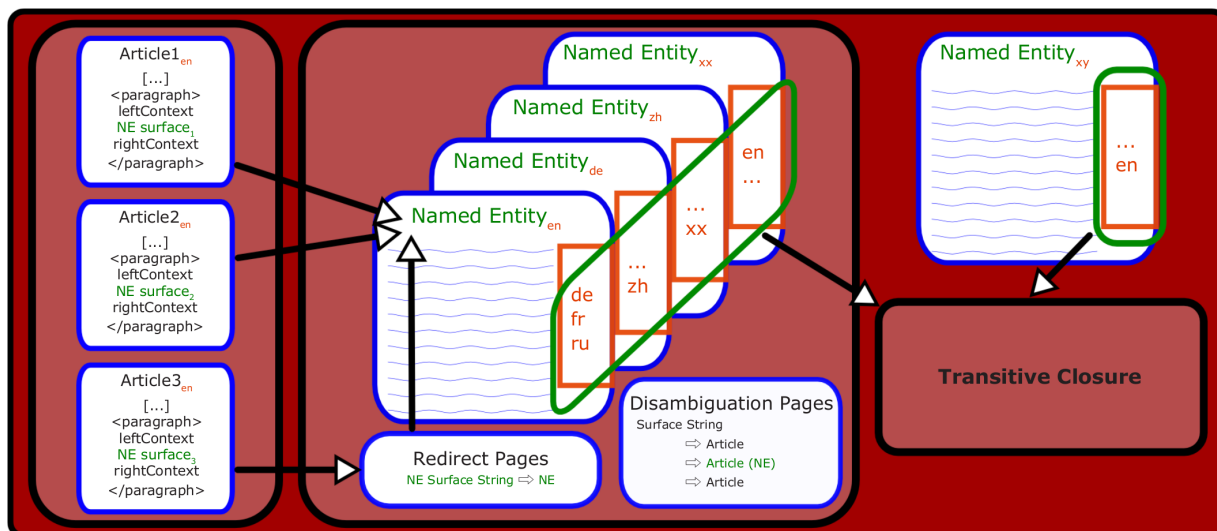


Abbildung 2.6: Erstellung des Named-Entity-Wörterbuchs

ist, ob ein Wikipediaartikel von einer Named Entity handelt, haben wir die Übereinstimmungsrate der Annotierer nach Cohens Kappa (erster Annotationssatz) bzw Fleiss' Kappa (zweiter Annotationssatz) berechnet (Cohen, 1960; Fleiss, 1971).

Tabelle 2.2 zeigt die Ergebnisse. Die Übereinstimmungsraten von 0,774 auf dem ersten und 0,771 auf dem zweiten Annotiersatz zeigen, dass die Aufgabe nicht trivial ist, wobei der Wert im letzteren Fall vor allem daher rührt, dass ein Annotierer von der Meinung der beiden anderen abwich. Insgesamt erreicht die Heuristik von (Bunescu and Pasca, 2006) eine Präzision von 95% und findet 1.547.586 Named Entities in der englischen Wikipedia³, womit wir eine sehr verlässliche Grundlage haben, um den Übersetzungsschritt durchzuführen.

Berechnung der transitiven Hülle (Triangulierung) über Sprachlinks

Wie bereits beschrieben lassen sich zu Wikipedia-Artikeln leicht Übersetzungen finden, indem man den Sprachlinks folgt und die Liste der Übersetzungen durch Triangulierung verbessern. Tabelle 2.3 zeigt die in der Publikation veröffentlichten Werte für den Zuwachs an Übersetzungen in den 13 größten Sprachen der Wikipedia sowie Türkisch und Swahili, die sich durch Triangulation berechnen lassen. Insgesamt wurden 77.694 (+4,47%) neue Übersetzungen entdeckt.

Allerdings beruhen diese Zahlen auf nur einer Iteration über die XML-Datenbankdumps der 16 Sprachen. Da diese sehr zeitaufwändig sind, konnte erst später ein weiterer Durchlauf folgen, der die Anzahl der Übersetzungen auf 147.797 (+8,89%) verdoppelte.

³XML Datenbankdump enwiki-20080103-pages-articles.xml

Überblick Annotationsdaten			
Menge	Kandidaten	Annotierer	Kappa
1	2000	2	0,774
2	2000	3	0,771

Ergebnis erster Annotationssatz			
Annotierer	True Positives	False Positives	Precision
1	1900	100	0,950
2	1924	76	0,962

Ergebnis zweiter Annotationssatz			
Annotierer	True Positives	False Positives	Precision
1	1872	128	0,936
2	1911	89	0,956
3	1914	86	0,957

Paarweise Annotiererübereinstimmung			
Annotierer	1	2	3
1	1,000	0,734	0,922
2	0,734	1,000	0,684

Tabelle 2.2: Evaluierungsergebnisse für Named Entity Erkennung

Der beachtliche Zuwachs deutet darauf hin, dass unsere Methode sinnvoll ist, um die per Hand gesetzten Sprachlinks zu vervollständigen.

Mit der geplanten Erweiterung auf MYSQL-Datenbanken wird die Berechnung voraussichtlich deutlich effizienter werden, sodass weitere Iterationen und das Einbinden von allen Wikipedia-Sprachversionen zu einem merklichen Anstieg der Übersetzung führen sollte.

Extrahierte Kontexte

Die für jede disambiguierte Named Entity gesammelten Kontexte haben im Englischen einen Umfang von mehr als 43 Millionen (bei ca. 1,5 Millionen Named Entities), wobei es ein paar wenige Named Entities gibt, denen besonders viele Kontexte zugeordnet sind, weil sie besonders oft innerhalb der Wikipedia verlinkt werden. Die meisten Kontexte waren den Named Entities *United States* (371.706) gefolgt von *England* (122.924), *United Kingdom* (114.140) und *Germany* (100.717) zugeordnet. Rund 1,5% aller Named Entities sind 44% der englischen Kontexte zugeordnet.

Dennoch haben die Kontexte von weniger häufig verlinkten Named Entities computerlinguistisch relevanten Umfang. Die Kontextfunde in anderen Sprachen sind entsprechend der Anzahl gefundener Übersetzungen und Wikipediagrößen niedriger.

Sprache	Initial	Final	Differenz	
			Absolut	%
de	243.903	250.049	6.146	2,46
es	127.518	137.606	10.088	7,33
fi	67.095	71.052	3.957	5,57
fr	215.479	222.712	7.233	3,25
it	135.852	145.889	10.037	6,88
ja	116.488	120.056	3.568	2,97
nl	166.708	176.203	9.495	5,39
no	63.431	66.786	3.355	5,02
pl	128.078	134.250	6.172	4,60
pt	132.778	137.227	4.449	3,24
ru	81.331	87.224	5.893	6,76
sv	97.270	99.710	2.440	2,45
sw	2.765	2.962	197	6,65
tr	26.814	29.059	2.245	7,73
zh	56.652	59.071	2.419	4,10
Gesamt	1.662.162	1.739.856	77.694	4,47

Tabelle 2.3: Verbesserung der Übersetzungszahlen durch die Triangulierungsmethode. Die *Initial*-Spalte enthält die Anzahl der Named Entities, die in der Zielsprache durch Folgen der Sprachlinks der Englischen Wikipedia gefunden werden. Die *Final*-Spalte zeigt die Anzahl der Named Entities nach der Triangulation.

2.2.4 Ausarbeitung der Publikation

Die ersten Ergebnisse der Erweiterung des Softwareprojekts waren so vielversprechend, dass wir unser Projekt bei der Language Resources and Evaluation Conference (LREC) 2008⁴ eingereicht haben. Da wir zwar mit dem Verfassen von wissenschaftlichen Arbeiten vertraut waren, eine Publikation jedoch etwas mehr Erfahrung benötigt, unterstützte uns Matthias Hartung dabei, die Ergebnisse von *HeiNER* schriftlich festzuhalten. Auch Frau Professor Dr. Frank stand uns wie auch zuvor schon mit vielen Vorschlägen und Ideen zur Seite.

Unsere Arbeit mit dem Titel *Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration* wurde angenommen und ist unter http://heiner.cl.uni-heidelberg.de/pub/HeiNER_lrec2008.pdf abrufbar.

2.2.5 Erstellen des Posters

Die Ergebnisse unserer Arbeit haben wir bei einer Postersession auf der LREC 2008 in Marrakesch vorgestellt. Da wir zum ersten Mal mit dem DIN A0 Format eines Posters umgehen mussten, gab es zunächst Probleme mit Dateiformaten und den Umsetzungsmöglichkeiten. Letztendlich entschieden wir uns dafür, die Gestaltung mit einer Testversion von Adobe Illustrator vorzunehmen, während die große Übersichtsgra-

⁴<http://www.lrec-conf.org/lrec2008/>

fik (siehe Figur 2.6) mithilfe der freien Programme *Inkscape*⁵ und *gimp*⁶ entstand. Die Herausforderung bestand darin die Ideen, Werkzeuge und Ergebnisse von *HeiNER* übersichtlich auf dem Poster zusammenzufassen, was zu zahlreichen Entwurfs-ideen und -änderungen führte. Das Ergebnis der aufwändigen Arbeit ist unter http://heiner.cl.uni-heidelberg.de/pub/HeiNER_poster.pdf zu finden.

2.2.6 LREC 2008

Die Konferenz war für uns sehr aufschlussreich. Die zahlreichen Vorträge zu verschiedenen computerlinguistischen Themen rund um das Thema Sprachressourcen und deren Evaluierung brachten viele Erkenntnisse und interessante Anregungen mit sich. Unsere Postersession wurde von den Besuchern sehr wohlwollend aufgenommen. Eine Teilnehmer lobte die Arbeit mit den Worten „Quite impressive“ und eine Doktorandin der Forschungsgruppe des European Media Lab in Heidelberg⁷ war überrascht, in welcher kurzer Zeit *HeiNER* entstanden ist.

Im Anschluss an die LREC haben sich verschiedene Forschergruppen weltweit für *HeiNER* interessiert und die Ressource heruntergeladen. Unter anderem meldete sich Oren Etzioni vom Turing Center an der Washingtoner Universität, der das System *PanDictionary* (Sammer and Soderland, 2007) voraussichtlich mithilfe unserer Daten erweitern möchte.

Zusammenfassung und Ausblick

Im Zuge der Förderung des Karl-Steinbuch-Stipendiums ist *HeiNER*, eine mehrsprachige Named Entity Ressource mit mehr als 1,5 Millionen Englischen Named Entities sowie deren Übersetzungen und einer Sammlung von disambiguierten Kontexten, entstanden. Das Ergebnis stellten wir auf der Language Resources and Evaluation Conference 2008 in Marrakesch vor, dessen Besuch uns erst durch das Stipendium ermöglicht wurde. Auch nach Auslauf des Stipendiums wollen wir *HeiNER* weiter verbessern, ausbauen und der Wissenschaft frei zur Verfügung stellen. Zum Beispiel muss die ursprünglich geplante API-Erweiterung zum Zugriff auf MYSQL-Datenbankdumps noch implementiert werden und eine weitere Aufgabe ist es, den Named Entities semantische Klassen zuzuordnen, was wahrscheinlich im Zuge einer Magisterarbeit geschehen wird.

Danksagung

Wir danken Anette Frank, Matthias Hartung, Nils Reiter und Philipp Cimiano für zahlreiche wertvolle und hilfreiche Diskussionen und Hinweise während der Arbeit an *HeiNER*. Außerdem danken wir der MFG-Stiftung Baden-Württemberg für die Auszeichnung der Arbeit mit dem Karl-Steinbuch-Stipendium 2008.

⁵www.inkscape.org

⁶<http://www.gimp.org/>

⁷<http://www.eml-research.de/>

Literaturverzeichnis

- Sisay Fissaha Adafre and Maarten de Rijke. Finding similar sentences across multiple languages in Wikipedia. In *Proceedings of the EACL 2006 Workshop on New Text-Wikis and Blogs and other Dynamic Text Sources*, 2006.
- I. Alegria, N. Ezeiza, and I. Fernandez. Named entities translation based on comparable corpora. In *Proceedings of the EACL Workshop on Multi-Word-Expressions in a Multilingual Context*, Genoa, Italy, 2006.
- Razvan Bunescu and Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, Trento, Italy, pages 9–16, April 2006.
- Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.
- Christiane Fellbaum. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998.
- Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.
- Rada Mihalcea. Using Wikipedia for automatic Word Sense Disambiguation. In *Proceedings of the NAACL 2007*, Rochester, April 2007.
- Simone Paolo Ponzetto and Michael Strube. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI-07)*, pages 1440–1447, July 2007.
- Marcus Sammer and Stephen Soderland. Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons. In *In Proceedings of the Machine Translation Summit XI 2007*, 2007.
- Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the 7th Conference on Natural language Learning at HLT-NAACL 2003*, pages 142–147, Morristown, NJ, USA, 2003.
- Richard Sproat, Tao Tao, and ChengXiang Zhai. Named entity transliteration with comparable corpora. In *Proceedings of the 44th Annual Meeting of the ACL*, Sydney, 2006.