Notizen zur 1. Sitzung "Parallele Korpora"

Sebastian Pado

19. Oktober 2010

Notizen zur 1. Sitzung des Seminars "Parallele Korpora". Ohne Gewähr.

- Linguistisch: CL ist Erforschung der Regelmässigkeiten im formalen System Sprache mit dem Computer bzw. formalen Modellen als Werkzeugen: Grammatikalität, Typologie etc.
- Kognitiv: CL ist Erforschung (insbesondere Modellierung) der Eigenschaften menschlichen Sprachverstehens: Lesezeiten, Plausibilitätsurteile, Referenz
- Ingenieurwissenschaftlich: CL ist Konstruktion sprachverarbeitender Systeme

Diese Veranstaltung wird primär die dritte Perspektive einnehmen; manchmal werden wir auch die erste Perspektive streifen (zB im Hinblick auf Übersetzung).

1 Was ist Sprachverarbeitung?

Was ist das Ziel von Sprachverarbeitung?

- Dass ein Computer sich "richtig verhält" (z.B. bei Informationszugriffsausgaben relevante Information zurückliefert, in einem Dialogsystem eine relevante Antwort gibt, etc.)
- Dazu muss der Computer in irgendeiner Art und Weise "verstehen", was der Benutzer eingibt (bzw. was Dokumente besagen)
- Wir nehmen hier an, dass es dafür notwendig ist, von der sprachlichen Oberfläche eine Art *Bedeutungsrepräsentation* (welcher Art auch immer) abzuleiten

Warum ist die Errechnung von Bedeutungsrepräsentationen schwierig?

- Sprachverarbeitung findet daher typischerweise in Form einer Verarbeitungspipeline statt, deren einzelne Schritte jeweils eine *weitere Beschreibungsebene explizit machen* – gegeben alle bisher explizierten Beschreibungsebenen
- In Sprache ist sehr viel Information *implizit* (Fachterminus: "hidden variables")

  - In gesprochener Sprache ist die Segmentierung nicht explizit
  - Die Wörter müssen mit der Struktur kombiniert werden, um die Bedeutung zu errechnen (Kompositionalität)
  - Die Dialogintention ergibt sich oft nicht direkt aus der wörtlichen Bedeutung
- Die einzelnen Schritte sind oft nicht deterministisch (Ambiguität!)
  - Wörter können verschiedene Wortarten haben
  - Wörter können verschiedene Bedeutungen haben
  - Sätze können verschiedene Strukturen haben

2 Explikation von Beschreibungsebenen: Klassifikation

- Typischerweise geht es bei den Explikationsschritten darum, Eingabeeinheiten Ausgabekategorien aus einer Menge möglicher Ausgabekategorien zuzuweisen: *Annotation*
  - Z.B. Phonem(sequenzen) in der Spracherkennung
  - Z.B. Wortarten beim POS-Tagging, etc.
- Es handelt sich also im technischen Sinne um *Klassifikation*
  - Der Klassifikator enthält damit explizites oder implizites *linguistisches Wissen*
  - Kann symbolisch/regelbasiert stattfinden, oder statistisch
  - Statistisch ist im Moment das dominierende Paradigma
    - NB. Parsing ist in dieser Hinsicht ein hybrider Klassifikationsprozess: fast alle Parsingmodelle sind statistisch, bauen aber auf symbolischen Generalisierungen (Syntaxregeln) auf

3 Überwachte und unüberwachte Klassifikation

- Klassifikation kann auf zwei Arten stattfinden
  - Überwachte (supervised) Klassifikation: Gegeben ein annotierter Datensatz (mit Eingabeeinheiten und Ausgabekategorien), verwende überwachte Lernverfahren, um einen Klassifikator zu trainieren, der anderen Eingabeeinheiten entsprechende Ausgabekategorien zuweisen kann
    - Beispiele: Support Vector Machines, Maximum Entropy Models, etc.
    - Bei sprachlichen Daten müssen die annotierten Korpora im Allgemeinen gross sein!
  - Unüberwachte (unsupervised) Klassifikation: Gegeben ein unannotierter Datensatz (nur Eingabeeinheiten), verwende distributionale Eigenschaft, um selbsttätig Kategorien aufzustellen un den Eingabeeinheiten zuzuweisen

    - Beispiele: Clustering
- Wenn man für unüberwachte Annotation keine annotierten Korpora braucht, warum verwenden dann nicht alle unüberwachte Verfahren?
    - Antwort 1: Weil es schwer ist, unüberwachte Verfahren dazu zu bringen, *linguistisch sinnvolle* Kategorien zu erzeugen
    - Antwort 2: Weil die Wahl der Features, die in unüberwachten Verfahren verwendet wird, grossen Einfluss auf die induzierten Kategorien hat, und es keine guten Regeln gibt, um "sinnvolle" Features zu wählen: die Schwierigkeit wird also gewissermassen nur verschoben.
- Beispiel 1: Wortartenzuweisung. Bei der unüberwachten Induktion von Wortarten für das Englische (Goldwater 2008) wurde das Wort "pea" (Erbse) fälschlicherweise in dieselbe Kategorie gesteckt, in der sich auch viele Adjektive befanden. Wieso? Weil in dem Korpus "pea" fast nur in der Konstruktion "pea soup" vorkam: eine klassische Adjektivposition.
- Beispiel 2: Grammatikinduktion. Für das folgende Dreisatz"korpus"

```
Peter sees a man.
Peter laughs.
Peter meets a friend.
```

  kann eine unüberwachte Grammatikinduktion zB folgende Regeln liefern:

```
S -> X Y
X -> N V
Y -> e
Y -> a N
```

  Diese Regeln beschreiben die Strukturen im Korpus, entsprechen aber nicht den gängigen Konventionen von Phrasenstrukturgrammatiken (Kombination Subjekt+Verb in Konstituente X; Konstituente Y kann durch "nichts" realisiert werden.

4 Projektionsverfahren

Die Schlussfolgerung aus diesen Beobachtungen ist, dass (zumindest im Moment) kein Weg um überwachte Lernverfahren herumführt. Wenn man aber nicht die Ressourcen hat, um monolingual Annotation zu erzeugen, kann man versuchen, parallele Korpora in einem von zwei Szenarien einzusetzen:

- Annotationsprojektion
- Oberflächenmerksmalsprojektion

Sei die "Zielsprache" die Sprache, in der wir Wissen induzierne möchten. Sie die "Quellsprache" die andere Sprache.

4.1 Annotationprojektion

Verfahren ist einsetzbar, wenn die quellsprachliche Seite des parallelen Korpora bereits annotiert ist bzw. automatisch annotiert werden kann

- Quellsprache annotieren (mit Klassifikator für Quellsprache, oder vielleicht existiert schon Annotation für die Quellsprache)
- Alinierung der Einheiten der Quell- und Zielsprache
- Kopie der Annotation von den quellsprachlichen auf die zielsprachlichen Einheiten
- (Falls nötig:) Filtern
- (Falls nötig:) Trainieren eines Klassifikators für die Zielsprache, der andere Korpora in der Zielsprache annotieren kann

4.2 Oberflächenmerkmalsprojektion

Verfahren ist einsetzbar, wenn die zielsprachliche Annotation von Interesse durch *Oberflächenmerkmale* der Quellsprache ausgedrückt wird. Beispiele: Wortbedeutungen (lassen sich durch Übersetzungen approximieren); Grammatische Funktionen im Englischen (lassen sich durch morphologische Markierungen im Deutschen approximieren); aspektuelle Eigenschaften von europäischen Sprachen (lassen sich durch Morphologie des Arabischen approximieren) etc.

- Definition von Äquivalenzklassen über quellsprachlichen Oberflächenmerkmalen
- Alinierung der Einheiten der Quell- und Zielsprache
- Annotation von zielsprachlichen Einheiten mit der jeweiligen Äquivalenzklasse
- (Falls nötig:) Filtern
- (Falls nötig:) Trainieren eines Klassifikators für die Zielsprache, der andere Korpora in der Zielsprache annotieren kann

Verhältnis zur Charakterisierung oben: Andere Sprache macht die verdeckte Information ("hidden variables") explizit

5 Annahmen von Projektionsverfahren

Projektionsverfahren machen eine Reihe von Annahmen, mit deren Grad an Erfüllung die Anwendbarkeit der Verfahren steht und fällt. Im allgemeinen müssen diese Annahmen neu evaluiert werden

- für jede Art von Annotation, die projiziert werden soll
- für jedes Sprachpaar

5.1 Übertragbarkeit von Kategorien

(Trifft insbesondere für Annotationsprojektion zu)

- Annahme 1: Die Kategorien der Quellsprache müssen auf die Zielsprache anwendbar sein – d.h. sinnvolle linguistische Kategorien der Zielsprache beschreiben.
- Das ist selbst für verwandte Sprachen nicht immer gegeben. Und selbst wenn die Kategorien existieren, gibt es häufig ein Kontinuum zwischen "Die Kategorien existieren" und "Die Kategorien sind relevant". Beispiel: englische verbale Wortarten

  ```
  VB Verb Base Form
  VBD Verb Past Tense
  VBG Verb Gerund
  VBN Verb Past Participle
  VBP Verb non-3rd person sg pres
  VBZ Verb 3rd person sg pres
  ```

  Bei der Übertragung auf das Deutsche sind diese Kategorien (fast) alle noch existent, aber viele von ihnen nicht besonders sinnvoll. Es fällt auf, dass die englischen Kategorien offensichtlich anhand morphologischer Kriterien entworfen sind: cf. die Unterscheidung zwischen 3. Person Singular und nicht-3.Person sowie die Spezialkategorie für Gerundien.

Mit etwas Glück sind die Kategorien zumindest nicht überlappend – das heisst, dass für jede Kategorie aus der Quellsprache gilt, das sie n Kategorien aus der Zielsprache komplett abdeckt, oder umgekehrt. In diesem Fall lassen sich zumindest Äquivalenzklassen definieren. Falls die Klassen aber überlappen, geht das auch nicht mehr (einfach).

5.2 Alinierung

- Annahme 2: Die sprachlichen Einheiten von quell- und zielsprachlichen Sätzen im parallelen Korpus müssen aliniert sein
- Diese Annahme lässt sich wiederum in zwei Teile aufteilen: Die Alinierbarkeit und die tatsächliche Alinierung
- Alinierbarkeit: Haben alle sprachlichen Einheiten auf der einen Seite genau eine Entsprechung auf der anderen Seite?
  - Unalinierte Einheiten in der Zielsprache erhalten keine projizierte Information, wie bei: *declare opened* / **für** *eröffnet erklären*.
  - Mit mehreren Einheiten in der Quellsprache alinierte zielsprachliche Einheiten können konfligierende Information erhalten, wie bei *of-PREP the-DET weather-NN des-PREP/DET?? Wetters-NN*.
- Tatsächliche Alinierung: Echte Alinierungsverfahren können nicht alle gewünschten Alinierungen erzeugen (siehe nächste Vorlesung)

5.3 Korrespondenz

- Annahme 3: Wenn zwei sprachliche Einheiten Übersetzungen voneinander sind, erhalten sie dieselbe sprachliche Annotation ("direkte Korrespondenzannahme" - Hwa et al., 2002).
- Das heisst, die Übersetzung eines Nomens ist ein Nomen, die Übersetzung eines Subjektes ist ein Subjekt, etc. pp.
- Hier setzen wir die Anwendbarkeit der Kategorien bereits voraus
- Das stimmt natürlich nicht (immer).
- Zu welchem Grad die Korrespondenzannahme zutrifft, werden wir im Laufe der Veranstaltung für verschiedene Szenarien anschauen.

5.4 Repräsentativität

- Annahme 4: Der zielsprachliche Teil des parallelen Korpus muss repräsentativ für die Zielsprache sein
- Sonst schlägt die Generalisierung von auf dem Parallelkorpus induzierten Regelmässigkeiten auf den Rest der Zielsprache fehl
- Wieso ist das potentiell ein Problem? Parallele Korpora entstehen typischerweise durch Übersetzung. Wenn das zielsprachliche Material also aus einer anderen Sprache übersetzt ist, ist es unter Umständen typologisch von der Ursprungssprache beeinflusst: "shining through"
  - Verwendung von grammatischen Konstruktionen, Passiv, Wortstellung, ...
- Hier ergibt sich (insbesondere für Annotationsprojektion) ein Spannungsfeld: Im Hinblick auf Annahme 3 hätte man gerne ein wörtliche Übersetzung, sodass der "Abstand" zwischen Quell- und Zielseite des Korpus möglichst gering ist; im Hinblick auf Annahme 4 möchte man aber eine Übersetzung, die möglichst typisch für die Zielsprache ist. Je grösser der typologische Unterschied von Quell- und Zielsprache, desto grösser diese Spannung.

6 Vier Themenkomplexe

Auf der Webseite sind die Referatsthemen in fünf Themenkomplexe eingeordnet. Wie verhalten sich diese Komplexe zu dem eben gesagten?

- Der Komplex "Alignment" befasst sich, wenig überraschend, mit Annahme 2: der Herstellung von Korrespondenzen zwischen den Einheiten in den zwei Sprachen
- Der Komplex "Übersetzung und Parallelismus" befasst sich mit Annahmen 1 und 3 (die in der Praxis oft nicht wirklich unterschieden werden)

- Die Komplexe III und IV (Induktion von Wissen aus annotierten bzw. unannotierten Quellkorpora) entsprechen genau dem, was oben als Annotationsprojektion und Oberflächenmerkmalsprojektion bezeichnet ist
- Der Komplex "vergleichbare und nichtparallele Korpora" befasst sich mit der Frage, was man tut, wenn es keine parallelen Korpora gibt, und welche Strategien unter welchen Umständen trotzdem anwendbar sind.
  - Relevanz: parallele Korpora gibt es nur für bestimmte Domänen bzw. Texttypen, und bei weitem nicht für alle Sprachpaare!