Vorlesungsnotizen: Literaturrecherche

Sebastian Pado

3. November 2010

Diese Vorlesungsnotizen geben Handreichungen zu drei zentralen Aspekten des "wissenschaftlichen Lesens" (und sind damit, zumindest in indirekt, auch für das wissenschaftliche Schreiben interessant).

1 Das Ziel

Allein in einem kleinen Feld wie der Computerlinguistik werden jedes Jahr mehrere hundert Artikel veroeffentlicht. Dazu kommen relevante Artikel aus angrenzenden Fachbereichen. Zu jedem gegebenen Thema gibt es also so viel Material, dass man mit der Lektüre Wochen füllen könnte. Was also nötig ist, ist eine **Auswahl** – und diese muss durch ein **Ziel** motiviert sein. Was ist das übergreifende Ziel des "wissenschaftlichen" Lesens?

Das Ziel ist, eine **konzeptuelle Landkarte eines wissenschaftliches Teilfeldes** zu erwerben, die es einem erlaubt

- neues Wissen zu altem in Verbindung zu setzen
- die Ergebnisse von Studien zu bewerten ("was wurde gelöst?")
- drängende Fragen zu identifizieren ("was wurde nicht gelöst?")
- um ggf. selbst wissenschaftlich tätig zu werden

Das Lesen ist dazu nur **eine** Methode, die unter Umständen alleine auch nicht ausreicht: z.B. bei technischen Themen (Grammatikformalismen, Semantikkonstruktion, maschinellen Lernverfahren) ist es oft nötig, zu "üben": Beispiele durchzurechnen etc.

1.1 Konzeptualisierung und Komplexität

- Ein zentraler Bestandteil des wissenschaftlichen Fortschritts ist Konzeptualiserung: die abstrakte Charakterisierung von Problemen/Problemklassen
- Unser Feld ist ziemlich neu
- Die meisten Probleme, die wir heute beherrschen, sind konzeptuell ziemlich einfach (auch wenn sie oft technisch involviert sind)

- *Fast immer* lässt sich der konzeptuelle Anteil eines Problems auf zwei Sätze herunterbrechen
- Wissenschaftlich betrachtet, ist es sehr wichtig, zuerst den konzeptuellen Anteil des Problems zu verstehen
- Ist die Konzeptualisierung angemessen?
  - Bei schlechter Konzeptualisierung ist die Arbeit an den nachfolgenden technischen Fragen typischerweise umsonst
    * Beispiel: Semantische Netzwerke: in den 70ern/80ern dominant in der Computerlinguistik, heutzutage so gut wie verschwunden

2 Drei primäre Fragen

- Was für relevante Literatur gibt es und wie komme ich heran?
- Wie identifiziere ich relevante Literatur?
- Woraufhin lese ich?

3 Relevante Literatur

3.1 Was gibt es für Typen von Literatur?

Drei Haupttypen von Texten in der CL:

- Konferenzpapiere (6-8 Seiten)
- Zeitschriftenartikel (15-40 Seiten)
- Handbuchkapitel/Lehrbuchkapitel/Sammelbandartikel

Entsprechen dem Zyklus des Wissenschaftsbetriebs:

- Jährliche Konferenzen (Publikation binnen 6 Monaten)
- Fortlaufende Zeitschriften (Publikation binnen 1 Jahr)
- Handbücher, Lehrbücher alle paar Jahre

Entsprechen einer Spezialitätshierarchie:

- Konferenzen: geschrieben fuer Spezialisten in einem Feld, die bereits mit der Materie vertraut sind
- Journals: geschrieben fuer Forscher in verwandten Gebieten

- Lehr/Handbücher: geschrieben fuer CL insgesamt (Lehrbücher: für Studenten, Handbücher: für Wissenschaftler)

Umfang:

- Konferenz: 1 kleineres Ergebnis von lokaler Relevanz
- Journal: 1 grösseres Ergebnis bzw. eine Gruppe kleinerer Ergebnisse: muss globale Relevanz demonstrieren

Relativ neu in unserem Feld: Überblicksartikel in Zeitschriten, die keinen eigenen wissenschaftlichen Beitrag leisten, sondern einen Überblick über die Literatur bieten.

CL gehört zu den Konferenzdisziplinen - das hat sie mit der Informatik gemeinsam. Die Linguistik ist hingegen primär eine Zeitschriftendisziplin.

3.2 Konferenzen

(Jede dieser Konferenzen lässt sich problemlos googeln)

- Die renommierten "klassischen" CL-Konferenzen: ACL, EACL, NAACL
- Neuere, aber ebenso renommierte Konferenzen: COLING, EMNLP
- Spezifische Konferenzen, die Unterfelder abdecken: LREC, IWCS, MOL
- Weniger renommierte Konferenzen: Cicling, RANLP, IJCNLP
- Informatik-Konferenzen mit CL-Anteil: AAAI, IJCAI (KI), NIPS, ICML (ML), SIGIR, VLDB (Anwendungen)
- Kognitionswissenschaftlich: CogSci, AmLap
- Deutsche Konferenzen: Konvens, GSCL

3.3 Journals

- Die beste CL-Zeitschrift: Computational Linguistics
- Eine eher ingenieurwissenschaftliche ausgerichtete Zeitschrift: Natural Language Engineering
- "Second tier": Research in Language and Computation, ACM Transactions on Speech and Language Processing, ...
- Spezifischer: Language Resources and Evaluation, Machine Translation
- Linguistisch: Int. Journal of Corpus Linguistics, Corpus Ling. and Linguistic Theory
- Informatik: AI Journal, Machine Learning
- Deutsch: Journal of Language Technology

3.4 Zugang

Konferenzen:

- $(NA|E|\epsilon)$ACL: über www.aclweb.org frei verfügbar
- Viele andere Konferenzen: über die ACM Digital Library
- Manche Konferenzen machen ihre Proceedings kostenpflichtig. In diesem Fall hat man die beste Chance über die Homepage des Autors bzw. der Autoren.

Zeitschriften:

- Computational Linguistics ist open access
- Viele Zeitschriften sind über die UB frei verfügbar
- Fast alle Zeitschriften erlauben den Autoren, "preprints" auf ihre Homepages zu stellen

Bei alten Artikeln hilft oft nur noch der Gang in die Bibliothek. Alternative: Dokumentenlieferservice Subito (5 Euro pro Artikel)

4 Identifikation relevanter Literatur

Wie identifiziere ich relevante Literatur?

- Googeln (aber das ist keine substantielle Antwort!)
- Citations (aber das ist auch nur ein ungefährer Anhaltspunkt)
- "Multiplikatoren": Referenzen in Lehrbüchern, Übersichtsartikeln
  - Referenzen in Tutorials für methodologische Themen
  - Referenzen in Papieren, die Sie schon kennen
- Achtung: Es ist eine schlechte Idee, alle Referenzen des ersten Papiers zu lesen, und dann wiederum alle Referenzen dieser Papiere
  - Am Ende landet man immer bei Chomsky oder Markov oder Frege

4.1 Die Struktur von Wissenschaft

(Bzw.: Wie ist die wissenschaftliche Publikationswelt strukturiert?)

- Forschung organisiert sich in "Clustern"
  - Alle Leute, die LFG machen, zitieren LFG; alle HPSG-Forscher zitieren HPSG
  - Gruppierung nach Phaenomen, Methode, Formalismus, ...

- Das erste Ziel einer Literaturrecherche ist es, *diese Cluster zu identifizieren*
  - D.h. man muss am Anfang "breiter" lesen, um einen Ueberblick zu bekommen, was fuer "Cluster" im Gebiet gibt
  - Dann sollte man sich aber auf den (oder die) relevantesten Cluster beschraenken und "tiefer" lesen, um die Diskussionen zu verstehen
- Aber es gilt immer noch: ein Papier zu lesen lohnt sich nur dann, wenn zu erwarten ist, dass es einen im Verständnis des Themas substantiell weiterbringt.
  - Dieser Fortschritt kann entweder konzeptuell sein oder technisch (eher selten, aber kommt vor)

5 Wie lese ich? Woraufhin lese ich?

Mit dem Lesen eines Papiers sollten drei konkrete Erkenntnisse verbunden sein.

- Erkenntnis 1: Ist das Papier relevant für mich?
- Erkenntnis 2: Was ist der Beitrag des Papiers?
- Erkenntnis 3: Was ist das *Problem* des Papiers – Wo knallt es?

Diese Erkenntnisse lassen sich anhand verschiedener Teile eines Papiers gewinnen - daher möchte ich als nächstes kurz über die Struktur von Papieren sprechen.

5.1 Formale Struktur eines Papiers

Konferenzpapiere stellen normalerweise eine **vierschrittige Argumentation** dar:

- Identifikation eines Problems und Motivation der Dringlichkeit des Problems ("Leidensdruck")
- Vorschlag einer Verbesserung (welcher Art auch immer)
- Evaluation der Verbesserung
- Rückbezug der positiven Evaluation auf das Problem

5.2 Inhaltliche Struktur eines Papiers

Bestandteile eines Papiers:

- Abstract: Das "Papier im kleinen". Typischerweise 3-5 Sätze (60-150 Wörter). Ein guter Abstract definiert das Problem und den Leidensdruck **im ersten Satz** und gibt Überblick ueber die Methoden, die zu einer Verbesserung geführt haben, und Schlüsselergebnisse.
- Einführung / Problemstellung (warum ist das interessant, was wir hier machen?)

- Lösungsvorschlag: neues System bzw. Modell
- Beweis bzw. Experimente: Detailebene
- Schluss: zurueck auf die allgemeine Ebene: Rückbezug auf das Problem, nächste Schritte, offene Probleme
- Related work": Wie verhält sich der Ansatz dieses Papiers zu anderen Papieren?
  - Position in einem Papier: offenes Problem.
  - Manche Leute: nach Einfuehrung ("das gibts und wir machen es anders")
  - Andere Leute: vor Schluss ("wir haben X gemacht und das ist anders als das, was es schon gibt")

Diese Struktur ist für Konferenzpapiere quasi unveränderlich, und auch die meisten Zeitschriftenpapiere halten sich daran.

5.3 Erkenntnis 1: Relevanz

- Relevanz sollte relativ einfach zu erkennen sein – zumindest, wenn der Erkenntnisgewinn konzeptuell sein soll (vgl. vorherigen Abschnitt). Im Prinzip sollte ein Blick auf den Abstract ausreichen.
- Es ist etwas anderes, wenn ich ein Papier auf seine technischen Verdienste hin durchlese, muss man es vermutlich tief lesen (zumindest die hinteren Abschnitte).

5.4 Erkenntnis 2: Beitrag

- Der Beitrag eines Papiers kann im Abstract aus Platzgründen oft nur kurz angerissen werden (manche Autoren nennen das Ergebnis auch absichtlich nicht in ihrem Abstract).
- Der eigentliche Beitrag ist im Allgemeinen in der Einführung des Papiers beschrieben, die oft oft ähnliches Material wie der Abstract abdeckt, aber noch einmal größeres Gewicht auf die genaue Problemdefinition legt.
- NB: Papiere sind im Allgemeinen für eine spezifische Zielgruppe geschrieben ("Fast-Spezialisten"). Unter Umständen ist der Beitrag eines Papiers in der Terminologie (s.u.) oder in einem Nebensatz "versteckt".
- Um ein Gefühl dafür zu kriegen, was der Beitrag eines Papiers ist, kann man außerdem den "related work"-Abschnitt lesen: dort wird im Idealfall explizit genannt, was das Papier von früheren Arbeiten im gleichen Feld unterscheidet.

5.5 Erkenntnis 3: Grenzen des Papiers

- Es ist wichtig, den Beitrag eines Papiers zu verstehen. Es ist mindestens genau so wichtig, die Grenzen eines Papiers zu verstehen.
  - Nicht zuletzt, weil ein Problem/eine Einschränkung eines Konferenzpapiers ein guter Einstieg für eigene Forschung sein kann... ;)
- Das ist aber oft schwieriger, denn oft wird liegt der Fokus von Papieren auf den positiven Ergebnissen und nicht so sehr auf den Grenzen.
- Das ist auch natürlich: Autoren wollen, dass ihre Ergebnisse veröffentlicht werden.
- Aber auch wenn das Argument überzeugend und/oder selbstbewusst vorgetragen wird, ist es deswegen noch lange nicht richtig.

Mehr Details im nächsten Abschnitt.

6 Wissenschaftskritik

6.1 Großkategorien

- **Relevanz**: Das Papier betrachtet ein Problem, das von geringem Interesse ist
- **Originalität**: Das Papier verwendet ein Verfahren, das bereits für dieses spezifische Problem verwendet worden ist
- **Motivation**: Das Papier motiviert seine Innovation nicht oder unzureichend
- **Technische Fehler / Ungenauigkeiten**: es sind Fehler in den Beweisen (theoretische Papiere) / Analysen (empirische Papiere) bzw. es legt wichtige Details seiner Methoden bzw. Evaluationsverfahren nicht offen
- **Übertreibung**: Das Papier behauptet mehr, als es zeigt

6.2 Details: Relevanz

- Ist das Problem wirklich relevant? Oder ist es vielleicht überhaupt kein Problem?
- "straw man": ein Problem, das nur aufgebaut wird, um es effektiv lösen zu können, bzw. eine Hypothese, die aufgebaut wird, um sie widerlegen zu können Beispiel: "Es wird oft behauptet, dass Deutsch schwieriger ist als Englisch. Wir zeigen anhand von Deutschen, die zwei Jahre Englisch gelernt haben, und Engländern, die zwei Jahre Deutsch gelernt haben, dass ihr Abschneiden in Tests nicht unterscheidbar ist."
  (Sowohl eine zweifelhafte Ausgangshypothese als auch ein zweifelhafter Schluss von einem Experiment auf die Ausgangshypothese)

6.3 Details: Motivation

- Ist der Leidensdruck überzeugend?
- Ist die vorgeschlagene Methode überzeugend?
  - Auch wenn sie das vorgestellte Problem löst, hat sie vielleicht andere Nachteile?
  - Beispiel: "wir lösen Probleme von POS-Taggern, indem wir parsen."
- Gäbe es vielleicht eine einfachere Lösung als die vorgeschlagene?

6.4 Details: Technische Fehler / Ungenauigkeit

Vielfach wird in Papieren auf die genaue Beschreibung des Verfahrens wenig Gewicht gelegt: es wird aus ausreichend angesehen, dass die Ergebnisse gut sind, und ggf. die Software verfügbar ist. Idealerweise sollte ein Papier **replizierbar** sein: Anhand der Informationen im Papier sollte das Verfahren nachprogrammierbar sein und man sollte per Evaluation die gleichen Ergebnisse erhalten können. Man kann also ein Papier daraufhin lesen, ob das möglich ist. Ein paar wichtige Informationen sind z.B. die folgenden:

- Wahl von freien Parametern bei statistischen Verfahren
- Genaue Beschreibung von Datenquelle, Sampling, Training/Test-Aufteilung, ...
- Vorverarbeitungsschritte
- "Normalisierung"
- Algorithmus eindeutig beschrieben?
- etc.

Ein zweites großes Problem bei der aktuellen, vorwiegend datengetriebenen, Forschung ist **Evaluation**: wir zeigen typischerweise gute Performanz auf einem bestimmten Testset und behaupten dann, dass die Ergebnisse auch auf andere Anwendungen übertragbar wären. Das ist nur unter ganz bestimmten Voraussetzungen der Fall. Die kann man die folgenden Fragen stellen:

- Sind die Daten repräsentativ?
  - Wenn das test set so gewählt ist, dass die vorgeschlagene Methode besonders gut funktioniert, sagt das Ergebnis gar nichts aus.
- Misst die gewählte Metrik wirklich Erfolg für das Problem?
  - Precision vs. Recall
- Was bedeuten die vorgestellten Zahlen? Handelt es sich um ein Niveau an Qualität, das interessant ist?

  - 20% Akkuratheit für eine neue Methode mag ein Erfolg sein, wenn die baseline bei 5% liegt – das heisst aber immer noch, dass 80% aller Antworten falsch sind

- Ist die Evaluation solide durchgeführt?
  - Signifikanztests?

6.5 Details: Übertreibung

- Von der Evaluation zur Conclusion machen Papiere gerne einen grossen Sprung – von spezifischen Evaluationsergebnissen hin zu allgemeinen Behauptungen.
- Generalisiert die Methode über das im Papier vorgestellte Szenario hinaus, oder ist sie auf das konkrete Problem zugeschnitten?
  - Viele Papiere stellen sich als "Pilotstudien" dar
  - Man sieht oft Strategien, die im Rahmen eines Papers gut funktionieren, aber die nie auf eine interessante Menge an Daten anwendbar wären
  - Mögliche Probleme: großer Aufwand in der Erstellung von Ressourcen; Abdeckungsprobleme; Effizienzprobleme; ...
- Werden auch negative Ergebnisse genannt, oder werden sie verschwiegen?
  - Manchmal gibt es subtile Hinweise im Design von Experimenten, die darauf hinweisen, dass nur bestimmte Ergebnisse berichtet werden sollen.

6.6 Exkurs: Darstellung vs. Inhalt

Hier geht es im inhaltliche Stringenz, nicht um die Darstellung: die ist oft auch suboptimal, aber das ist für einen (eigentlich) Leser nur ein zusätzliches Problem, das einen nicht von der inhaltlichen Ebene ablenken sollte. Manchmal (insbesondere bei Autoren mit anderen Muttersprachen) ist die Darstellung aber so schlecht, dass es sehr schwer ist, insbesondere zwischen sprachlichen Unzulänglichkeiten und Ungenauigkeiten in der Argumentation zu unterscheiden.

6.7 Extra-Erkenntnisse: Terminologie

- In jedem Forschungsfeld gibt es Begriffe, die auf ganz bestimmte Art und Weise verwendet werden (Fachvokabular).
- In einem guten Papier werden die wichtigen Konzepte früh eingeführt und benannt, d.h. im Abstract bzw. der Einleitung.
- Das gilt u.U. nicht für Konzepte, die im Papier selbst entwickelt werden: die muss man dann im Modellierungs/Systemabschnitt suchen.

7 Am Ende des Lesens

- Nach dem Lesen sollten Sie in der Lage sein, einen Text zu formulieren, der den Inhalt des Papiers in **einem kurzen** Abschnitt in ihre eigenen Worten zusammenfasst (ähnlich der Einleitung in einem Konferenzpapier, ohne den "Papierstruktur"-Abschnitt).

- Wenn Ihnen das nicht im Rahmen einer Seite gelingt, handelt es sich entweder um ein außerordentlich komplexes Problem, oder Sie haben die zugrundeliegende Konzeptualisierung noch nicht genau genug erkannt. Lesen Sie das Papier noch einmal - und wenn es schwer zu verstehen ist (z.B. weil es viel voraussetzt), lesen Sie zuerst das Vorgängerpapier, auf das sich das aktuelle Papier am intensivsten stützt.

- Außerdem sollten Sie skizzieren können, was Sie für das gravierendste Problem bzw. die gravierendste Einschränkung des aktuellen Papiers halten. Auch das sollten Sie in einem Abschnitt zusammenfassen können.