Tutorial Notes

Educational Natural Language Processing

AIED 2009, Brighton, July 7

Iryna Gurevych, Delphine Bernhard and Aljoscha Burchardt
Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt

Presenters

Iryna Gurevych (gurevych@tk.informatik.tu-darmstadt.de)
Iryna Gurevych is head of the Ubiquitous Knowledge Processing (UKP) Lab at the University of Darmstadt. Her recent research has focused on the application of lexical semantic knowledge in such areas as spoken dialogue summarization, information retrieval for educational purposes, e.g. electronic career guidance, or question answering based on question-answer repositories in Web 2.0 applied to eLearning. Her areas of expertise include algorithms for computational lexical semantics and processing of user generated discourse. She guided the development of the high-performance Java-based Wikipedia and Wiktionary APIs as well as projects in collaborative annotation, information filtering and sentiment analysis for eLearning.

Delphine Bernhard (delphine@tk.informatik.tu-darmstadt.de)
Delphine Bernhard is Senior Researcher in the Ubiquitous Knowledge Processing (UKP) Lab at the University of Darmstadt. She obtained her PhD in 2006 from the Université de Grenoble 1, where she worked on terminology extraction from domain specific texts and unsupervised morphological analysis. Her current work focuses on enhancing question answering systems to meet the specific needs of learners. Her further research topics include processing user generated discourse and quality assessment of social media content.

Aljoscha Burchardt (burchardt@tk.informatik.tu-darmstadt.de)
Aljoscha Burchardt is scientific coordinator of the Center of Research Excellence “eLearning 2.0” and Senior Researcher in the Ubiquitous Knowledge Processing Lab at the University of Darmstadt. He obtained his PhD from Saarland University in 2008, where he worked in projects related to both eLearning and applied lexical semantics. His current work focuses on the use of summarization techniques to access and present multimodal learning materials in collaborative settings.

Overview

Typical Web 2.0 tools such as wikis, blogs, and podcasts have recently entered the classroom and foster interactions between learners and tutors, within the new eLearning 2.0 paradigm. As a result, eLearning 2.0 makes large amounts of eLearning discourse available for Natural Language Processing (NLP) within the field of research that we call "Educational Natural Language Processing" (e-NLP). Research on e-NLP has existed for a long time and has focused on e.g. intelligent tutoring systems (Litman & Forbes-Riley, 2006), or essay scoring (Attali & Burstein, 2006). This field of research brings together two communities: language technology on the one side and educational computing on the other side. Several workshops on "Building Educational Applications Using NLP" and related topics have already taken place at major conferences, such as HLT-NAACL 2003, COLING 2004, ACL 2005, ACL 2008 and NAACL-HLT 2009.

NLP techniques are used in many educational applications working with textual data such as intelligent tutoring systems or computer-assisted language learning. However, these applications are particularly challenging for NLP since they require an adaptation of NLP techniques to various types of discourse, e.g. tutoring dialogues, which are different from typical task-oriented spoken dialogue systems. Moreover, educational applications place strong requirements on NLP systems, which have to be robust yet accurate. Therefore, this is an important application domain and a source of innovation for both NLP and educational computing, as shown by Feng et al. (2006), Kim et al. (2006), Malioutov & Barzilay (2006) and Csomai & Mihalcea (2007), to name just a few.

In this tutorial, we will review a variety of uses of NLP in the educational domain and point to emerging trends which call for new types of applications.

Contents **Slide numbers**

1. Introduction: eLearning and NLP....7-12
2. Automatic generation of exercises....14-43
   a) Computer-based testing....15-19
   b) Multiple-choice questions....20-23
   c) Fill-in-the-blank questions....24-27
   d) Multiple-choice cloze questions....28-32
   e) Matching test items....33-35
   f) Error correction questions....36
   g) Evaluation....37-43
3. Assessment of learner generated discourse....44-102
   a) Introduction....44-50
   b) Assessing short textual answers....51-64
   c) Essay grading....65-83
   d) Plagiarism....84-102
4. Reading and writing assistance....103-144
   a) Text readability....103-109
   b) Document retrieval for reading practice....110-113
   c) Text simplification....114-115
   d) Vocabulary assistance....116-127
   e) Spell checking....128-134
   f) Grammar checking....135-140
   g) Dictionary lookup....141-144
5. Web 2.0 and computer supported collaborative learning....145-171
   a) Web 2.0 & eLearning 2.0....145-153
   b) NLP for Wikis....154-169
   c) Quality of user-generated discourse....170-171
6. Example e-NLP application 1: electronic career guidance....172-187
7. Example e-NLP application 2: educational question answering....188-203
8. Conclusions....204-210
9. Bibliography....Appendix

Educational Natural Language Processing (ENLP)

Delphine Bernhard, Aljoscha Burchardt,
Iryna Gurevych

TECHNISCHE UNIVERSITÄT DARMSTADT

The Presenters

TECHNISCHE UNIVERSITÄT DARMSTADT


Iryna Gurevych


Delphine Bernhard

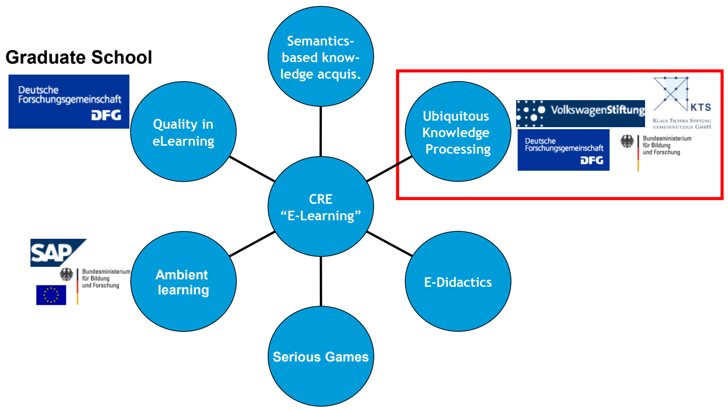

Aljoscha Burchardt

Technische Universität Darmstadt

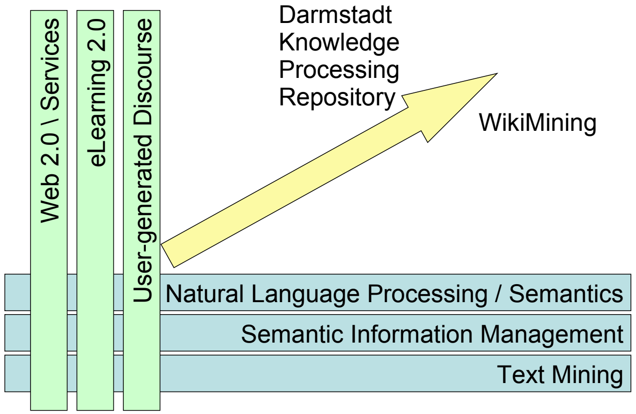
TECHNISCHE UNIVERSITÄT DARMSTADT

Center of Research Excellence eLearning 2.0

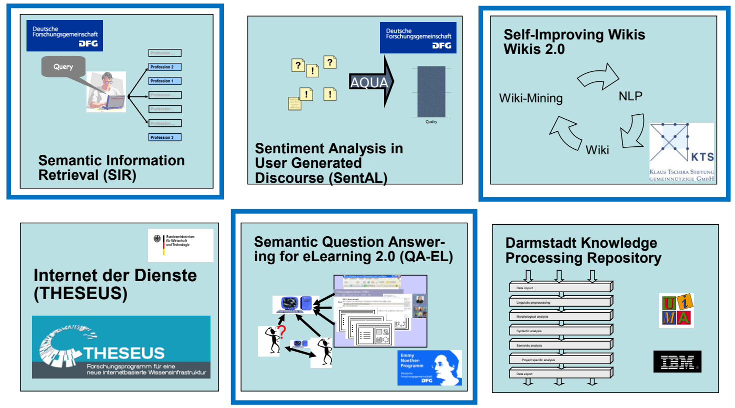
TECHNISCHE UNIVERSITÄT DARMSTADT



UKP Lab Research Topics



Research Projects and eLearning



Outline

- **Introduction: eLearning and NLP**
- Automatic generation of exercises
- Assessment of learner generated discourse
- Reading and writing assistance
- Web 2.0 and computer supported collaborative learning
- Example e-NLP application: electronic career guidance
- Example e-NLP application: educational QA
- Wrap up and questions

e-NLP

Educational Natural Language Processing

eLearning → Computer-assisted learning / instruction

NLP → Analysis and use of language by machines

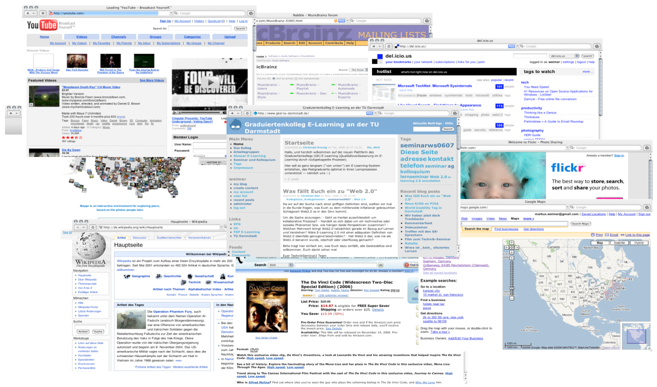
Definition

TECHNISCHE UNIVERSITÄT DARMSTADT

Field of research exploring the use of NLP techniques in educational contexts

Web 2.0 & eLearning 2.0

TECHNISCHE UNIVERSITÄT DARMSTADT



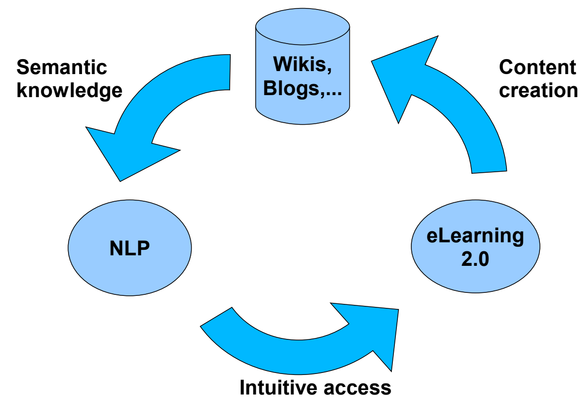
Some Observations

TECHNISCHE UNIVERSITÄT DARMSTADT

- Large text repositories with user generated discourse and user generated metadata are created
- These repositories need advanced information management and NLP to be efficiently accessed
- Using these repositories to create structured knowledge bases can improve NLP

Feedback Loop: NLP & eLearning 2.0

TECHNISCHE UNIVERSITÄT DARMSTADT



Outline

Introduction: eLearning and NLP
Automatic generation of exercises
Assessment of learner generated discourse
Reading and writing assistance
Web 2.0 and computer supported collaborative learning
Example e-NLP application: electronic career guidance
Example e-NLP application: educational QA
Wrap up and questions

Outline

Introduction: eLearning and NLP
Automatic generation of exercises
Assessment of learner generated discourse
Reading and writing assistance
Web 2.0 and computer supported collaborative learning
Example e-NLP application: electronic career guidance
Example e-NLP application: educational QA
Wrap up and questions

Computer-based Testing

- **Definition**: *All forms of assessment delivered with the help of computers*
- Also called:
  - Computer Assisted/Aided Assessment (CAA)
- Adequate question types for CAA (McKenna & Bull, 1999):
  - Multiple choice questions (MCQs)
  - True/False questions
  - Matching questions
  - Ranking questions
  - Sequencing questions
  - etc.

Question Types

- **Objective** test items
  - **constrained** answer, to be selected among a set of alternatives
  - **short** answer (word or phrase) in response to a question
  - **objective** and **impartial** scoring
- Examples:
  - Fill-in-the-blanks questions
  - Multiple-choice questions
  - Matching questions
- **Subjective** test items
  - **original** answer
  - **variable length**
  - **biased** scoring
- Examples:
  - Short-answer essays
  - Extended-response essays

Roles of Test Items in Learning

TECHNISCHE UNIVERSITAT DARMSTADT

- **Summative assessment**
  - "Assessment **of** learning"
  - Measuring student achievement
- **Formative assessment**
  - "Assessment **for** learning"
  - Active learning: encourage learners to practice and apply newly acquired knowledge by answering test items

NLP for CAA

TECHNISCHE UNIVERSITAT DARMSTADT

- **Generation of questions and exercises**
  - Writing test questions, especially objective test items, is an extremely difficult and time consuming task for teachers
  - Use of NLP to automatically generate **objective** test items, esp. for language learning
- **Assessment and evaluation of answers to subjective test items**
  - Use of NLP to automatically:
    - Diagnose errors in short-answer essays
    - Grade essays

Automatic Generation of Test Items

TECHNISCHE UNIVERSITAT DARMSTADT

- **Source data**
  - Corpora: texts should be chosen according to
    - the learner model (level, mastered vocabulary)
    - the instructor model (target language, word category)
  - Lexical semantic resources, e.g. WordNet
- **Tools**
  - Tokeniser and sentence splitter
  - Lemmatiser
  - Conjugation and declension tools
  - POS tagger
  - Parser and chunker

Multiple-Choice Questions (MCQ)

TECHNISCHE UNIVERSITAT DARMSTADT

- Choose the correct answer among a set of possible answers
- Example (Mitkov et al., 2006)

Who was voted the best international footballer for 2004?

(a) Henry
(b) Beckham
(c) Ronaldinho
(d) Ronaldo

Distractors / Distracters

Key

Question / Stem

- Usually 3 to 5 alternative answers

Distractors

- **Distractors** (also **distracters**) are the incorrect answers presented as a choice in a multiple-choice test
- Generation of "**good**" distractors (McKenna & Bull, 1999; Duvall)
  - Ensure that there is only one correct response for single response MCQ
  - The key should not always occur at the same position in the list of answers
  - Distractors should be grammatically parallel with each other and approximately equal in length
  - Distractors should be plausible and attractive
  - However, distractors should not be too close to the correct answer and risk confusing students

Automatic Generation of MCQs

1. **Selection of the key**
   - Unknown words that appear in a reading (Heilman & Eskenazi, 2007)
   - Domain-specific terms:
     - Automatically extracted (Mitkov et al., 2006)
     - Present in a thesaurus, e.g. UMLS (Karamanis et al., 2006)

2. **Generation of the stem**
   - Constrained patterns (Heilman & Eskenazi, 2007):
     Which set of words are most related in meaning to "reject"?
   - Transformation of source clauses to stems, using transformation and agreement rules (Mitkov et al., 2006):
     Transitive verbs require objects → Which kind of verbs require objects?

Automatic Generation of MCQs

3. **Generation of the distractors**
   - WordNet concepts which are semantically close to the key, e.g. hypernyms and co-hyponyms (Mitkov et al., 2006; Karamanis et al., 2006)
     Stem: "Which part of speech serves as the most central element in a clause?"
     Key: "verb", Distractors: "noun", "adjective", "preposition"
   - Thesaurus-based and distributional similarity measures (Mitkov et al., 2006)
   - Other NPs with the same head as the key, retrieved from a corpus (Mitkov et al., 2006)
     Key: "transitive verbs", Distractors: "modal verbs", "phrasal verbs", "active verbs"

Fill-in-the-Blank Questions (FIB)

- Also called **cloze** test
- Technique which dates from 1953 (Wilson Taylor)
- Consists of a portion of text with certain words removed
- The student is asked to "fill in the blanks"
- **Objective cloze items** = multiple-choice cloze items, i.e. students are given a list of words to use in a cloze
- **Subjective cloze items** = students can choose the words
- Challenges:
  - Phrase the question so that only one correct answer is possible
  - Spelling errors in subjective cloze items

Fill-in-the-Blank Examples

- Blank = preposition (Source: http://www.purl.org/net/WERTI)

SANTIAGO , May 15 (Reuters) - Chile 's
Chaiten volcano groaned , rumbled and
shuddered **on** Thursday , raising new concerns
among authorities , [?] lightning bolts
pierced the huge clouds [?] hot ash
hovering ominously [?] its crater .

- Blank = verb to be conjugated (Source: http://www.nonstopenglish.com/exercise.asp?exid=915)

Fill in the gaps with the correct tenses: Past Simple or Present Perfect
Example: I (see already) _____ the Pope. (key = have already seen)

1. Yesterday she ______ (get) a new bed.
2. ______ (ever be) in London?
3. When was the last time you ______ (call) her?
4. What ______ (you do) when you saw her?

Fill-in-the-Blank Question Generation

1. Selection of an input corpus
2. POS tagging
3. Selection of the blanks in the input corpus
4. Where needed, provide some information about the word in the blank, e.g. verb lemma when the test targets verb conjugation (Aldabe et al., 2006)

Selection of the Blanks

- Every "n-th" (e.g. fifth or eighth) word in the text (Coniam, 1997)
- Words in specified frequency ranges, e.g. only high frequency or low frequency words (Coniam,1997)
- Words belonging to a given grammatical category (Coniam, 1997; Aldabe et al., 2006)
- Open-class words, given their POS, and possibly targeted word sense (Liu et al., 2005; Brown et al., 2005)
- Machine learning, based on a pool of input questions used as training data (Hoshino & Nakawaga, 2005)

Objective Multiple-Choice Cloze Items

Combination of a cloze item with multiple-choice answers

(adj) strange: *He thought it was .... that her mobile was switched off.*
- allegation
- sinister
- peculiar
- grieve
- virulent

(adj) strange: *He thought it was peculiar that her mobile was switched off.*
- allegation
- sinister
- ✓ peculiar
- grieve
- virulent

http://www.wordlearner.com

Generation of the Distractors

- Randomly chosen in the text from which the question was generated (Hoshino & Nakagawa, 2005)
- Same POS (Coniam, 1997)
- Similar frequency range (Coniam, 1997)
- For grammar questions, use a declension or a conjugation tool to generate different forms of the key, e.g. change case, number, person, mode, tense, etc. (Aldabe et al., 2006, Chen et al., 2006)
- Common student errors in the given context (Lee & Seneff, 2007)
- Collocations: frequent co-occurrence with either the left or the right context (Lee & Seneff, 2007)
- Open class words: semantic similarity based on distributional similarity (Smith et al., 2008) or a thesaurus (Sumita et al., 2005)

The Frequency Heuristic

(Coniam, 1997)

A University of Wollongong researcher, Ms. Robyn Iredale, commented that a __(2)__ of the hiring practices of 55 companies also said "there was no __(3)__ putting a small Asian in a __(4)__ of authority over taller Australians." She said: "They said __(5)__ workers would not like having Asians __(6)__ because they work too hard."

Table 4
Word Classes and Word Frequencies in Test Items

Item no.	Word (test key)	Word class tag	Frequency
2	survey	noun	1,715
3	point	noun	299
4	position	noun	632
5	other	determiner	80
6	around	preposition	201

Item (2)

	Option	Frequency
A.	driver	1,716
B.	distance	1,717
C.	survey [key]	1,715
D.	dream	1,719
E.	tree	1,724

Item (3)

	Option	Frequency
A.	war	210
B.	course	222
C.	point [key]	299
D.	lot	231
E.	thing	234

Verification of the Distractors

- Basic verifications:
  - there must be enough distractors
  - there must be no duplicated distractors (Aldabe et al., 2006)
- Collocations: choose distractors that do not collocate with important words in the target sentence (Liu et al., 2005; Smith et al., 2008)
- Use of the Web: if the sentence/phrase containing the distractor is frequent on the web, then the distractor should be rejected (Sumita et al., 2005)

The child's misery would move even the most ____ heart.

(a) torpid	hits("the most torpid heart") = 4	Good distractors because infrequent
(b) invidious	hits("the most invidious heart") = 0	
(c) stolid	hits("the most stolid heart") = 6	
(d) obdurate	hits("the most obdurate heart") = 1 240	

Student Project in the e-NLP Course at the TU Darmstadt

- Based on "Automatic generation of cloze items for prepositions" (Lee & Seneff, 2007)
- Example:
  If you don't have anything planned for this evening, let's go __ a movie.
  (a) to (b) of (c) on (d) null
- Tasks:
  - INPUT: sentence + key, OUTPUT: list of three distractors
  - The three distractors must each be generated taking a different approach
    - baseline: word frequencies
    - collocations
    - "creative" method, devised by the students
- Conclusion: a motivating and interesting project for students

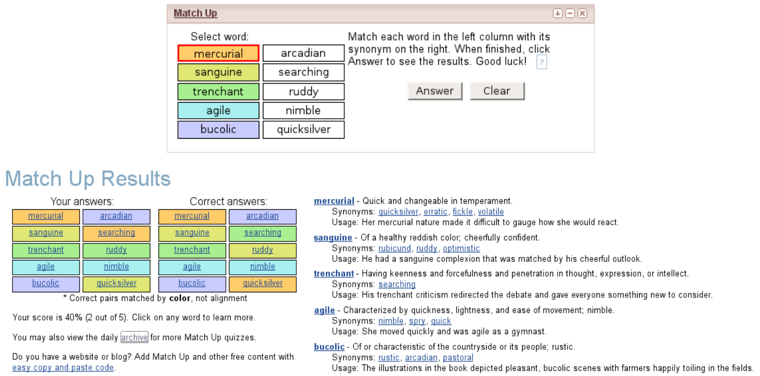
Matching Test Items

TECHNISCHE UNIVERSITÄT DARMSTADT

- Task: match items in one list with response items in another list
- Kinds of elements matched:
  - Word – synonym
  - Definition – term
  - Word – antonym
  - Hypernym – hyponym
  - Historical event – date
  - etc.
- Matching test items assess a learner's understanding of relationships

Matching Test Items

TECHNISCHE UNIVERSITÄT DARMSTADT



http://www.thefreedictionary.com

Matching Test Items for Vocabulary Assessment (Brown et al., 2005)

TECHNISCHE UNIVERSITÄT DARMSTADT

Wordbank:

verbose infallible obdurate opaque

Choose the word from the wordbank that best completes each phrase below:

1. ___ windows of the jail
2. the Catholic Church considers the Pope ___
3. ___ and ineffective instructional methods
4. the child's misery would move even the most ___ heart

Glosses for specific word senses in WordNet

Error Detection Questions

TECHNISCHE UNIVERSITÄT DARMSTADT

- Aim: detect and possibly correct errors, which can be marked or not
- Example (Chen et al., 2006)

  Although maple trees are <u>among</u> the most colorful varieties
  (A)
  <u>in the fall</u>, they lose $\boxed{\text{its}}$ leaves <u>sooner than</u> oak trees.
  (B) (C) (D)
- Wrong statements are produced by the distractor generator

Evaluation of Generated Questions

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Student evaluation**
  - Difficulty and response time
  - Comparison with results obtained for manually generated tests (Heilman & Eskenazi, 2007)
- **Instructor evaluation**
  - Usability: "all distractors result in an inappropriate sentence" (Liu et al., 2005; Lee & Seneff, 2007)
  - Post-editing: count how many test items are accepted, rejected or revised by instructors during post-editing (Aldabe et al., 2006; Mitkov et al., 2006)

Pre-requisites for Student Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

- **External assessment**
  - Evaluate the linguistic and / or factual knowledge of the students before they take the test , e.g. the Nelson-Denny Reading Test, the Raven's Matrices Test, the Lexical Knowledge Battery (Brown et al., 2005)
- **Self-assessment**
  - Have the students assess whether they know the target word or not (Brown et al., 2005; Heilman & Eskenazi, 2007)
    "Do you know the word 'w'?"

Item Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

- Investigate the quality of the test items (Zurawski, 1998)
- Quantitative item analysis:
  - **Facility / Difficulty index** (p): number of test takers who answered the item correctly divided by the total number of students who answered the item
  - **Discrimination index** (D): "does the test item differentiate those who did well on the exam overall from those who did not?"
    - Divide the students in two groups: high-scoring and low-scoring (above and below the median)
    - Compute the item difficulty index separately for both groups: p_{upper} and p_{lower}
    - Discrimination index $D = p_{upper} - p_{lower}$

Item Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

- Example

  The child's misery would move even the most ____ heart.

(a) torpid	chosen by 7 students
(b) invidious	chosen by 1 students
(c) stolid	chosen by 3 students
(d) obdurate	chosen by 15 students

  #Students: 26
- Difficulty index: 15 / 26 = 0.58 → neither too difficult nor too simple (recommended score: 0.5)
- Discrimination index
  - 9 out of 12 students in the high group found the correct answer
  - 6 out of 14 students in the low group found the correct answer
  - D = 9/12 – 6/14 = 0.75 – 0.43 = 0.32
  - The test item is a quite good discriminator

Item Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Item distractor analysis**: examine the percentage of students who select each incorrect alternative, to determine if the distractors are functioning well

Well-designed item

Possibly miskeyed

Candidate for removal

Candidate for revision

Distractor Analysis Data for Upper (*U*) and Lower (*L*) Scoring Students

	Item 1				Item 2				Item 3				Item 4			
	A	b	c	d	a	B	c	d	a	b	C	d	a	b	c	D
U	24	3	2	1	1	1	26	2	13	2	13	2	7	10	7	6
L	10	7	7	6	8	4	11	7	9	5	11	5	1	3	2	24

Note: Correctly keyed alternative for each item is identified in capitalized print.

Source: (Zurawski, 1998)

Efficiency of the Automatic Generation of Test Items

TECHNISCHE UNIVERSITÄT DARMSTADT

- Even though automatically generated test items have to be post-edited, this is still a lot faster than writing new test items from scratch
- Mitkov et al. (2006) report the following figures:
  - an average of 1 minute and 40 seconds was needed to post-edit a test item in order to produce a worthy item
  - an average of 6 minutes was needed to manually produce a test item

Summary

TECHNISCHE UNIVERSITÄT DARMSTADT

- The generation of questions and exercises is actually **semi-automatic**: the system's output has to be verified and modified by an instructor
- However, NLP-based systems considerably **reduce the time spent** by instructors to write test items, even if they have to manually correct the generated test items
- A great variety of **NLP technologies and resources** have been successfully used so far:
  - POS tagging and parsing
  - Word sense disambiguation
  - Term extraction
  - ...

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Introduction: eLearning and NLP
Automatic generation of exercises
Assessment of learner generated discourse
Reading and writing assistance
Web 2.0 and computer supported collaborative learning
Example e-NLP application: electronic career guidance
Example e-NLP application: educational QA
Wrap up and questions

Assessment of Learner Generated Discourse

TECHNISCHE UNIVERSITÄT DARMSTADT

- Discourse ≈ Utterance longer than a sentence
- Language **form**: written or spoken
- **Types** of learner generated discourse:
  - Emerging in institutional settings, e.g. solutions to exercises
  - Emerging in informal settings, e.g. discussions in forums (next section)

Importance of Institutional eAssessment

TECHNISCHE UNIVERSITÄT DARMSTADT

- Feedback to the student about her level of knowledge
- Feedback to the instructor about the progress of students' learning
- Incentive to study certain things, to study them in certain ways, to master certain skills
- Formal means for grading and/or making a pass/fail decisions

Importance of Free-Text Assessments

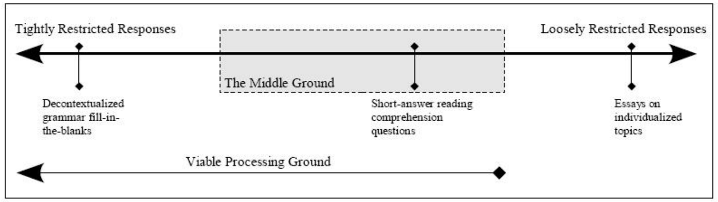
TECHNISCHE UNIVERSITÄT DARMSTADT

- Advantages over traditional multiple-choice assessments (Bennett & Ward, 1993)
- Major obstacle is the large cost and effort required for scoring
- Automatic systems:
  - Reduce these costs
  - Facilitate extended feedback to students

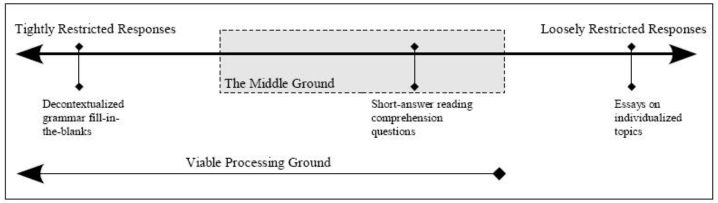
Learning Exercise Spectrum Model (Bailey & Meurers 2008)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Proposed in the context of language learning (ICALL), but applicable to different topics



Tasks Discussed in this Tutorial



MC-Tests
FIB

Assessing short textual answers

Essay grading

Detecting plagiarism

Relating Properties of the Tasks with NLP Techniques

Assessing short textual answers

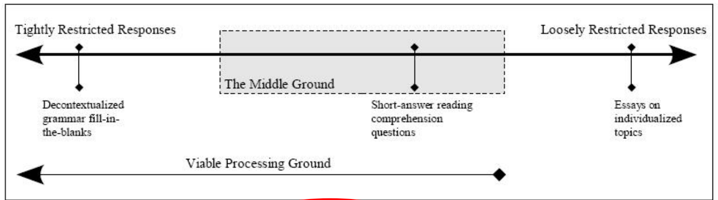
- Gold-standard answers can be provided
- Specific information must be complete and correct
- Word meaning (predicate-argument-structure) matters
- Ressource-based apprs.

Essay grading

(Detecting plagiarism)

- Unpredictable (no correct answer)
- Holistic (overall organiza-tion, style, etc.)
- Rhetorical structure matters
- Corpus-based approaches
- Supervised approaches

Tasks Discussed in this Tutorial



MC-Tests
FIB

Assessing short textual answers

Essay grading

Detecting plagiarism

Automatic Assessment

- Diagnosis, i.e., content assessment (CAM) on learner data
  - Language learning (Bailey and Meurers, 2008)
  - Error detection in C-rater (Leacock, 2004)
- Scoring of learner data (later)
  - Essays
  - Plagiarism
  - Speech

Detecting Meaning Errors (Bailey and Meuerers, 2008)

TECHNISCHE UNIVERSITAT DARMSTADT

- Analysis of responses to short-answer comprehension tests
  - 1-3 sentences in length
- Error codes:
  - Necessary concepts left out of learner response
  - Response with extraneous, incorrect concepts
  - An incorrect blend/substitution (correct concept missing, incorrect one present)
  - Multiple incorrect concepts
- Human disagreement in 12%, eliminated from the evaluation data

CUE: *What are the methods of propaganda mentioned in the article?*

TARGET: *The methods include use of labels, visual images, and beautiful or famous people promoting the idea or product. Also used is linking the product to concepts that are admired or desired and to create the impression that everyone supports the product or idea.*

LEARNER RESPONSES:

- *A number of methods of propaganda are used in the media.*
- *Bositive or negative labels.*
- *Giving positive or negative labels. Using visual images. Having a beautiful or famous person to promote. Creating the impression that everyone supports the product or idea.*

Technology of CAM

TECHNISCHE UNIVERSITAT DARMSTADT

- Input:
  - Learner's response, one + target responses, question, source reading passage
  - Linguistic analysis: annotation, alignment, diagnosis

Annotation Task	Language Processing Tool
Sentence Detection, Tokenization, Lemmatization	MontyLingua (Liu, 2004)
Lemmatization	PC-KIMMO (Antworth, 1993)
Spell Checking	Edit distance (Levenshtein, 1966), SCOWL word list (Atkinson, 2004)
Part-of-speech Tagging	TreeTagger (Schmid, 1994)
Noun Phrase Chunking	CASS (Abney, 1997)
Lexical Relations	WordNet (Miller, 1995)
Similarity Scores	PMI-IR (Turney, 2001; Mihalcea et al., 2006)
Dependency Relations	Stanford Parser (Klein and Manning, 2003)

Source:
(Bailey & Meurers, 2008)

Spell Checking Example (from Leacock & Chodorow, 2003)

TECHNISCHE UNIVERSITAT DARMSTADT

- 67 different variants of *Reagan in about 9,000* responses. Below are all the spelling variants of *Reagan* that occurred more than once:

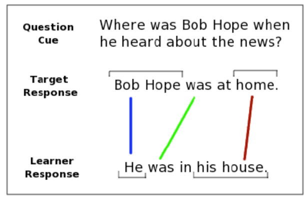
Regan, Reagon, Reagen, Raegan, Regans, Regean, Reagons, Ragan, Ragen,Reagin, Raegon, Regon, Reagn, Reagean, Reegan, Ragon, Ragean, Reagens,Raegen, Raegans, Reggan, Raygon, Rgan, Regens, Regen, Regeans, Reagion,Ragons, Raegin

- Spell checking not as easy a task as one would think
  - *Reagons* is close (in terms of edit distance) to the existing word *reasons*
  - Yet, in the domain of US presidents, *Reagan* is more probably the intended word

Technology of CAM

TECHNISCHE UNIVERSITAT DARMSTADT

- Alignment maps new concepts from learner's response to those in target
  - Token level (abstraction from string to lemma, semantic type (e.g. date, location)
  - *Houses* => *house* => LOC
  - Chunk level, e.g., *home* ≈ *his house*
  - Relation level (dependency, lexical)
  - Pronoun resolution



- Diagnosis analyzes if the learner's response contains content errors

Technology of CAM

TECHNISCHE UNIVERSITÄT DARMSTADT

- Given the alignment analysis, when is a learner input correct / faulty / wrong?
- Evaluation
  - Hand-written rules 81% on the development data, 63% on the test data
  - Machine learning (TiMBL), 88% accuracy on the test data for binary semantic error detection task
  - Viable results

C-Rater (Leacock & Chodorow, 2002)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Measures student understanding with little regard to writing skills
- Example question (4th grade math question used in the National Assessment for Educational Progress (NAEP)):

A radio station wanted to determine the most popular type of music among those in the listening range of the station. Would sampling opinions at a country music concert held in the listening area of the station be a good way to do this?

▢ YES ▢ NO

Explain your answer.

Technology of c-Rater

TECHNISCHE UNIVERSITÄT DARMSTADT

- Content expert develops a scoring guide
  - Gold standard responses
- Recognizing the equivalence of the response to the correct answers
  - Essentially paraphrase recognition
- Analysis in terms of:
  - regularizing over morphological variation
  - matching on synonyms or similar words
  - resolving the spelling of unrecognized words
  - resolving the referent of any pronouns in the response
  - predicate argument structure
- Mapping canonical representations to those of the gold standard responses
  - Rule-based

Predicate Argument Structure in c-rater

TECHNISCHE UNIVERSITÄT DARMSTADT

- Transform text to tuples (verbs and their arg.s): „atomic meaning units"

Score	Sentence and tuple		
Credit	Most people at the country show would say that country music is the most popular music.		
	say	**:subject most people**	
	be	**:subject country music**	**:object most popular music**
Credit	The people at the country concert would all answer country music.		
	answer	**:subject people**	**:object country music**
Credit	People at a country concert might think that country music is the best music.		
	think	**:subject people**	
	be	**:subject country music**	**:object best music**
No credit	I happen to like country music and so do most of my friends.		
	like	**:subject I**	**:object country music**
	do	**:object most of my friends**	

2003)

Problems with this Simple Approach to Predicate Argument Structure (Excursion)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Variation in language is much more pervasive
- Simple example: passive voice
  - *Mary ate the cake.* (subject: *Mary*)
  - *The cake has been eaten by Mary.* (subject: *the cake*)
- Simple solution: check for passive (syntactic parser) and switch arguments
- Harder example:
  - *John is afraid of Ghosts.*
  - *Ghosts scare John.*
- Solution: Use a semantic ressource like FrameNet.

Frame Semantics and FrameNet
(Fillmore 1976, Baker et. al. 1998)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Lexical semantic classification of predicates and their argument structure
- A frame represents a prototypical situation (e.g. Commercial_transaction, Theft, Awareness)
- A set of roles identifies the participants or propositions involved
- Frames are organized in a hierarchy
- Berkeley FrameNet Project db: 600 frames, 9.000 lexical units, 135.000 annotated sentences

Linguistic Normalization
(Frame: Commerce_buy)

TECHNISCHE UNIVERSITÄT DARMSTADT

Role	Example Sentence
Seller	*BMW bought Rover from British Aerospace.*
Buyer	*Rover was bought by BM ... ed [...] the new Range Rove*
Goods	*BMW, which acquired Rover in 1994, is now dismantling the company.*
Money	*BMW's purchase of Rove ... n was a good move.*

Voice: active / passive

Lexicalization

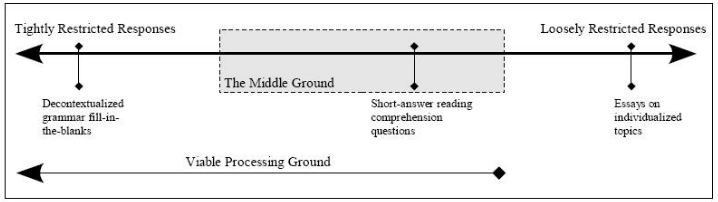
POS: verb / noun

Wrapping up Content Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

- Applicable for short, predictable answers
- Usually ressource-based
  - Spell-checkers, Grammars
  - Semantic ressources
  - Special rule-based systems
  - …
- A Result of a c-rater experiment (Leacock and Chodorow 2003)
  - About 84% agreement with human judgment
  - 47% baseline for majority class (full / partial / no credit)

Tasks Discussed in this Tutorial



MC-Tests
FIB

Assessing short textual answers

Essay grading

Detecting plagiarism

What is an Essay?

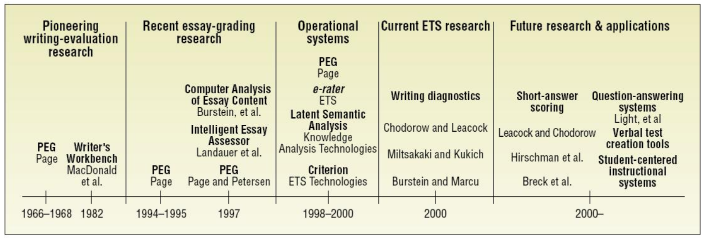
- A major part of formal education (at least in the USA)
- Secondary students are taught structured essay formats to improve their writing skills
- Often used by universities in selecting applicants
  - Students are asked to explain, comment on, or assess a topic of study
  - These admission essays are used to judge the mastery and comprehension of the material

Essay Prompts

- **Descriptive prompt:**
  - “Imagine that you have a pen pal from another country. Write a descriptive essay explaining how your school looks and sounds, and how your school makes you feel.”
- **Persuasive prompt:**
  - “Some people think the school year should be lengthened at the expense of vacations. What is your opinion? Give specific reasons to support your opinion.”

Source: Y. Attali and J. Burstein. Automated essay scoring with e-rater v.2. The Journal of Technology, Learning, and Assessment, 4(3), February 2006.

Research Development in Writing Evaluation



Source: Marti A. Hearst, The Debate on Automated Essay Grading, IEEE Intelligent Systems, IEEE Educational Activities Department, 2000, 15, 22-37.

Most Prominent Systems

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Intelligent Essay Assessor** (Landauer, Foltz & Laham, 1998)
  - Based on a statistical technique for summarizing the relations between words in a document, i.e. every word is a „mini-feature“
- **Intellimetric** (Elliot, 2001)
  - Based on hundreds of undisclosed features
- **Project Essay Grade –** PEG (Page, 1994)
  - Based on dozens of mostly undisclosed features
- **E-Rater** (Burstein et al., 1998)
  - The 1st version used more than 60 features
  - E-rater 2.0 uses a small set of features

How Do Humans and Machines Rate Essays?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Humans evaluate various **intrinsic variables** of interest → essay score:
  - Content adequacy
  - Structure
  - Argumentation
  - Diction
  - Fluency
  - Correct language use
- Machines use **approximations** or **possible correlates** of intrinsic variables → scoring model

How is a Scoring Model Created?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Analyze a few hundred essays:
  - Written on a specific prompt
  - Pre-scored by as many human raters as possible
- Identify most useful approximations (classification features) out of those available to the system
- Employ a statistical modeling procedure to combine the features and produce a machine-generated score

Validating the Meaning of Scores (Yang et al. 2002)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Relationship between human and machine scores of the same prompt:
  - Compare the machine-human and human-human agreement (Burstein et al., 1998; Elliot, 2001; Landauer et al., 2001)
  - Estimate a true score as the one assigned by multiple raters (Page, 1966)
- Relationship between test scores and other similar measures:
  - Compare automatic scores with multiple-choice test results and teacher judgments (Powers et al., 2002)
- Understanding the scoring process, i.e. relative importance of different writing dimensions:
  - Most commonly used features in scoring models (Burstein et al., 1998)
  - The most important component is content (Landauer et al., 2001)

Skepticism and Criticism (Page and Petersen, 1995)

- Three general directions of criticism:
  - **Humanistic** – never understand or appreciate an essay as a human

    → Use automatic scoring as a second rater
  - **Defensive** – playful or hostile students produce "bad faith" essays

    → a study by Powers et al. (2001), a lot of data needed
  - **Constructive** – computer-measured variables is not what is really important for an essay

    → an improved ability to additionally provide diagnostic feedback

Features Used by e-Rater 2.0 (Burstein et al., 1998)

- Measures of:
  - Grammar, usage, typos
  - Style
  - Organization & development
  - Lexical complexity
  - Prompt-specific vocabulary usage
- Implemented in different *writing analysis tools*
- Based on an NLP foundation that provides instructional feedback to students in the web-based *Criterion* system

Writing Analysis Tools: Correctness

- Identify five main types of grammar, usage and mechanics errors:
  - Agreement and verb formation errors, wrong word use, missing punctuation, typographical errors
- Corpus-based approach:
  - Train the system on a large corpus of edited text
  - Extract and count bigrams of words and POS
  - Search for bigrams in essay that occur much less often (Chodorow & Leacock, 2000)
  - *girl walk* occurs less frequently than *girl walks*

Writing Analysis Tools: Aspects of Style

- The writer may wish to revise:
  - The use of passive sentences
  - Very long or very short sentences
  - Overly repetitious words (Burstein & Wolska, 2003)

Writing Analysis Tools: Organization & Development

TECHNISCHE UNIVERSITÄT DARMSTADT

- Discourse elements present or absent in the essay (Burstein, Marcu and Knight, 2003)

- A linear representation of text as a sequence of:
  - Introductory material
  - A thesis statement
  - Main ideas
  - Supporting ideas
  - A conclusion

- How can we find these parts automatically ?

Supervised Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

- Train a system on a large corpus of human annotated essays to identify "good" sequences

- The computer extracts regularities such as
  - Mandatory parts,
  - Number restriction, e.g., > 3 main ideas,
  - …

Essay Annotated with Discourse Elements

TECHNISCHE UNIVERSITÄT DARMSTADT

<Introductory Material>“You can’t always do what you want to do!,” my mother said. She scolded me for doing what I thought was best for me. It is very difficult to do something that I do not want to do.**</Introductory Material>** **<Thesis>**But now that I am mature enough to take responsibility for my actions, I understand that many times in our lives we have to do what we should do. However, making important decisions, like determining your goal for the future, should be something that you want to do and enjoy doing.**</Thesis>**

<Introductory Material>I’ve seen many successful people who are doctors, artists, teachers, designers, etc.**</Introductory Material>** **<Main Point>**In my opinion they were considered successful people because they were able to find what they enjoy doing and worked hard for it.**</Main Point>** **<Irrelevant>**It is easy to determine that he/she is successful, not because it’s what others think, but because he/she have succeed in what he/she wanted to do.**</Irrelevant>**

<Introductory Material>In Korea, where I grew up, many parents seem to push their children into being doctors, lawyers, engineer etc.**</Introductory Material>** **<Main Point>**Parents believe that their kids should become what they believe is right for them, but most kids have their own choice and often doesn’t choose the same career as their parent’s.**</Main Point>** **<Support>**I’ve seen a doctor who wasn’t happy at all with her job because she thought that becoming doctor is what she should do. That person later had to switch her job to what she really wanted to do since she was a little girl, which was teaching.**</Support>**

<Conclusion>Parents might know what’s best for their own children in daily base, but deciding a long term goal for them should be one’s own decision of what he/she likes to do and want to do **</Conclusion>**

Source: Y. Attali and J. Burstein. Automated essay scoring with e-rater v.2. The Journal of Technology, Learning, and Assessment, 4(3), February 2006.

Writing Analysis Tools: Lexical Complexity

TECHNISCHE UNIVERSITÄT DARMSTADT

- Related to word-specific characteristics such as:
  - A measure of vocabulary-level, based on Breland, Jones and Jenkins (1994), *Standardized Frequency Index* across the words in an essay
  - The average word length in characters in an essay

Writing Analysis Tools: Prompt-Specific Vocabulary Usage

TECHNISCHE UNIVERSITÄT DARMSTADT

- Intuition: good essays resemble each other in their word choice, as will poor essays (within the same prompt)

- Idea: compare an essay to a sample of essays from each score category (usually 1-6)
  - Each essay and a set of training essays from each score category is converted to a vector
  - Some function words are removed
  - Each vector element is a weight based on a word frequency function
  - Six cosine correlations are computed between the essay and each score category to determine the similarity

Scoring in e-Rater 2.0

TECHNISCHE UNIVERSITÄT DARMSTADT

- Input: all features of all writing analysis tools
  - Grammar, usage, mechanics, style (4 features)
  - Organization & development (2 features)
  - Lexical complexity (2 features)
  - Prompt-specific vocabulary usage (2 features)

- Straightforward combination method:
  - Apply a linear transformation on feature values to achieve a desired scale
  - A weighted average of the standardized feature values

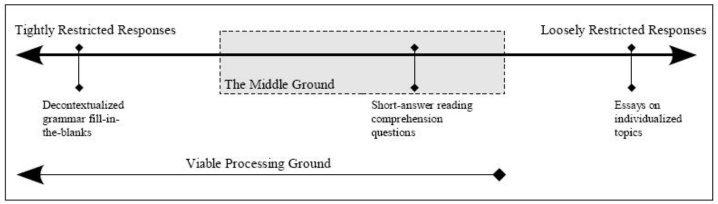
Future Directions

TECHNISCHE UNIVERSITÄT DARMSTADT

- Better standardization of scoring - a single scoring model for all prompts of a program or assessment
- Better understanding and control over the automated scores
- Cover more aspects of writing quality, devise new features
  - Prefer features providing useful instructional feedback
- Detection of anomalous and bad-faith essays
  - Characterize different types of anomalies
  - Detect off-topic essays (Higgins, Burstein and Attali, 2006)

Tasks Discussed in this Tutorial

TECHNISCHE UNIVERSITÄT DARMSTADT



MC-Tests FIB

Assessing short textual answers

Essay grading

Detecting plagiarism

Plagiarism

“Plagiarize: […] to take and use as one's own the thoughts, writings, or inventions of another. […]”

Oxford English Dictionary Online

- Main Feature: Missing indication of source

Affected Types of Media

- Music
- Text
- Graphics
- Images
- …

- In this context: written text

Plagiarism at Universities

- Two common kinds of plagiarism among students
  - Intra-corpal plagiarism
    - Copying from fellow students
    - Kollusion (here: unwanted group work)
  - Web-based plagiarism
    - Copying from an online source (book, web page, etc.)

(Culwin and Lancaster 2001)

- „Web 2.0-mentality“: *Find-Remix-Share*

(Sattler 2007)

Plagiarism at Universities (Lecturers/Researchers)

- In teaching material: slides / course reader / etc.

- Self-plagiarism
- Silent inclusion of results in one‘s own work (from PhD candidates, students, etc.)

(http://www.spiegel.de/unispiegel/jobundberuf/0,1518,207062,00.html)

- Peer-Reviews (project proposals, conference papers)

(http://de.wikipedia.org/wiki/Plagiat#Plagiate_in_Hochschule_und_Schule)

- Honorary authorship

Types of Plagiarism

TECHNISCHE UNIVERSITÄT DARMSTADT

(1) Plagiarism of authorship: the direct case of putting your own name to someone else's work

(2) Word-for-word plagiarism: copying of phrases or passages from a published text without quotation or acknowledgement.

(3) Paraphrasing plagiarism: words or syntax are changed (rewritten), but the source text can still be recognized.

(4) Plagiarism of the form of a source: the structure of an argument in a source is copied (verbatim or rewritten)

(5) Plagiarism of ideas: the reuse of an original thought from a source text without dependence on the words or form of the source

(6) Plagiarism of secondary sources: original sources are referenced or quoted, but obtained from a secondary source text without looking up the original.

Based on Martin (1994) and Clough (2003)

Typical Plagiarism Indicators

TECHNISCHE UNIVERSITÄT DARMSTADT

- Use of advanced or technical vocabulary beyond that expected of the writer
- A large improvement in writing style compared to previous submitted work
- Inconsistencies within the written text itself, e.g. changes in vocabulary, style (e.g. references) or quality
- Incoherent text where the flow is not consistent or smooth
- Dangling references: a reference appears in the text, but not in the bibliography and vice versa
- A large degree of similarity between the content, mistakes, etc. of two or more submitted texts.

Based on Clough (2003)

Techniques Used to Conceal Copying

TECHNISCHE UNIVERSITÄT DARMSTADT

- Replacing odd or unusual words
- Changing formatting
- Adding filler words or phrases
- Changing headings
- Rephrasing sentences
- Removing or re-ordering sections
- Changing spelling (usually from American English to British English, if the document is plagiari[s|z]ed from the Web)
- Producing consistency by find-and-replace (as an example, if some papers refer to the World Wide Web, some to the WWW, some to the Web, a student may perform a global find-and-replace to ensure consistency within the plagiarised document)
- In programming, changing variable names and comments

The use of electronic tools to support plagiarism detection: http://www.comp.leeds.ac.uk/hannah/CandIT/plagiarism.html

String Matching Algorithms

TECHNISCHE UNIVERSITÄT DARMSTADT

- Most popular **plagiarism detection scheme**:
  - Comparing word windows of length ≥ *n*
  - Computing the overlap of matching subsequences and substrings (consecutive tokens)
  - n is derived empirically
  - The longer n becomes, the more unlikely it is that the same sequence will appear in independently written texts
  - Problem: larger n-grams types are rare, difficult to define thresholds

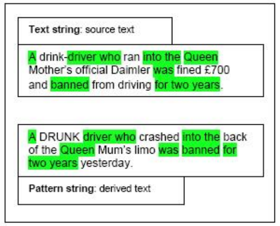
Uniqueness of N-grams
(from Clough 2003)

▪ Figures taken from 769 texts in the METER corpus:

N (words)	N-gram occurrences (tokens)	Distinct n-grams (types)	% distinct n-grams	% distinct n-grams in 1 file
1	137204	14407	11	39
2	248819	99682	40	67
3	248819	180674	73	82
4	257312	214119	85	90
5	251429	226369	90	93
6	250956	231800	92	94
7	250306	234600	94	95
8	249584	236310	95	96
9	248841	237409	95	97
10	289610	278903	96	97

Table 1 Uniqueness of consecutive n-word sequences (n-grams) as *n* increases from 1-10 words

Longest Common Substrings Computed between Two Sentences



▪ **Greedy String Tiling** (or *GST:* see, e.g. (Wise,1993)), an algorithm which computes a **maximal mapping** of text pairs with **non-overlapping substrings** (called *tiles)*.

▪ Advantage: n-gram size needs not be set *a priori*

Longest Common Substrings Computed between Two Sentences

▪ The output of the GST algorithm is a list like: [for two years], [driver who], [into the], [a], [queen], [was] and [banned].

▪ Different quantitative measures can then be applied, e.g.:
  ▪ the minimum and maximum tile length
  ▪ the average tile length
  ▪ the dispersion of tile lengths

▪ Goal: derive a similarity measure for plagiarism

▪ Challenge: distinguish derived and non-derived text(s)

Example of Tiling for Derived and Non-Derived Text (from Clough 2003)

▪ It has been empirically found that:

▪ derived texts (top) share longer matching substrings

▪ the tiling for a derived and non-derived text pair are in most cases apparently different

Machine Learning in Plagiarism Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

- Input: Documents and their features (Document length, match size, etc.)
- Goal: A computational model that distinguishes original and plagiarism
- **Supervised (machine) learning**: train a classifier on manually annotated training data (texts classified as plagiarized or not)
    - Disadvantage: Many documents needed (thousands)
- **Unsupervised learning:** have the machine find certain "clusters"
    - Concrete instruction: Divide these texts in two parts (given these features)
    - Hope: one part will contain originals and one part derived texts
- Evaluation: check random samples

Relaxing the Approach

TECHNISCHE UNIVERSITÄT DARMSTADT

Preserving longer matching n-grams and tile lengths to make the approach resistant to simple edits

- Allow small gaps to represent token deletion
- Allow simple word substitution (using WordNet)
- Allow insertion of certain words such as domain-specific terminology and function words (e.g. conjunctions)
- Allow simple reordering of tokens (e.g. transposition)

NLP in Plagiarism Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

- Existing work involves minimal natural language processing (NLP)
- Areas of NLP that could aid plagiarism detection, particularly in identifying texts which exhibit similarity in semantics, structure or discourse, but differ in lexical overlap and syntax
- NLP methods include:
    - morphological analysis, part-of-speech tagging, anaphora resolution, parsing (syntactic and semantic), co-reference resolution, word sense disambiguation, and discourse processing
- Future work:
    - several similarity scores based on lexical overlap, syntax, semantics, discourse and other structural features

How to Avoid Plagiarism?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Clearly define plagiarism to the students and use explicit examples
- Educate the students about the honor code and the ramifications if it is violated
- Create assignments that make plagiarism difficult
- Make sure the students are familiar with online resources
- Have the students submit evidence of the research process as well as the paper
- Avoid repeating assignments and paper topics
- Inform the students you are Internet savvy and you know about the paper mills (visit the sites with the students to evaluate the quality of the work)
- Inform the students that you use plagiarism detection software

From "Plagiarism in the 21st century" *Carrie Leslie. Lunch & Learn. 2004. Otto G. Richter Library*

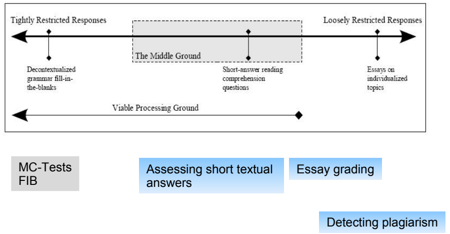
Online Internet Plagiarism Services

TECHNISCHE UNIVERSITÄT DARMSTADT

- Plagiarism.org www.plagiarism.org
  - The largest online plagiarism service available
- EVE2 www.canexus.com/eve/abouteve.shtml
- None of the services details their implementation details
- All of them are commercial, but plagiarism.org allows free trial

Summing up

TECHNISCHE UNIVERSITÄT DARMSTADT



- Resource-based vs. corpus-based approaches
- Resources: spell checker, grammar, thesaurus, semantic net, …
- Corpus-based approaches
  - Supervised: Manual annotation and generalization
  - Unsupervised: Automatic induction of structure

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

- Introduction: eLearning and NLP
- Automatic generation of exercises
- Assessment of learner generated discourse
- **Reading and writing assistance**
- Web 2.0 and computer supported collaborative learning
- Example e-NLP application: electronic career guidance
- Example e-NLP application: educational QA
- Wrap up and questions

Readability

TECHNISCHE UNIVERSITÄT DARMSTADT

- "Readability is what makes some texts easier to read than others" (DuBay, 2004)
- Heavily dependent on the intended audience
- A text's readability can be estimated with readability formulas, which provide an objective prediction of text difficulty, usually expressed in terms of school grade level
- Aims:
  - match reading materials with the abilities of the readers
  - support authors in writing clearly understandable texts

Traditional Readability Measures

TECHNISCHE UNIVERSITAT DARMSTADT

Formula	Date	Features	Example values
Flesch index	1948	- average # syllables / word - average sentence length	- 30 = "very difficult" - 70 = "easy"
Fog index	1952	- # words with more than 2 syllables - average sentence length	- 6 = comic books - 10 = newspapers
SMOG grading	1969	- # words with more than 3 syllables	- 0 to 6 = low-literate - 19+ = post-graduate

Readability Statistics

TECHNISCHE UNIVERSITAT DARMSTADT

- Computed using the `style` command

Rotkäppchen


```
readability grades:
        Kincaid: 7,0
        ARI: 6,5
        Coleman-Liau: 7,5
        Flesch Index: 77,7/100
        Fog Index: 8,7
        Lix: 25,5 = below school year 5
        SMOG-Grading: 2,2
sentence info:
        5406 characters
        1364 words, average length 3,96 characters = 1,31 syllables
        74 sentences, average length 18,4 words
        40% (30) short sentences (at most 13 words)
        20% (15) long sentences (at least 28 words)
        38 paragraphs, average length 1,9 sentences
        8% (6) questions
        24% (18) passive sentences
        longest sent 42 wds at sent 58; shortest sent 3 wds at sent 13
sentence beginnings:
        pronoun (8) interrogative pronoun (6) article (7)
```

DIE ZEIT

```
readability grades:
        Kincaid: 11,3
        ARI: 12,1
        Coleman-Liau: 16,3
        Flesch Index: 42,1/100
        Fog Index: 13,9
        Lix: 42,8 = school year 7
        SMOG-Grading: 7,5
sentence info:
        5336 characters
        980 words, average length 5,44 characters = 1,76 syllables
        62 sentences, average length 15,8 words
        45% (28) short sentences (at most 11 words)
        14% (9) long sentences (at least 26 words)
        9 paragraphs, average length 6,9 sentences
        0% (0) questions
        27% (17) passive sentences
        longest sent 48 wds at sent 13; shortest sent 2 wds at sent 17
sentence beginnings:
        pronoun (9) interrogative pronoun (0) article (9)
```


Statistical Language Models for Reading Difficulty

TECHNISCHE UNIVERSITAT DARMSTADT

- Use of statistical models representing norms, specific populations and individuals (Brown & Eskenazi, 2004)
- Different models can be created for each level of reading difficulty (Collins-Thompson & Callan, 2005)
- Method (Collins-Thompson & Callan, 2005; Heilman et al., 2007, 2008):
  - For a given text passage T, the semantic difficulty of T relative to a specific grade level G_i is predicted by calculating the likelihood that the words of T were generated from a representative language model of G_i
  - Reading difficulty = grade level of the language model most likely to have generated the passage T

Readability analysis as a classification task

TECHNISCHE UNIVERSITAT DARMSTADT

- Aim: label texts with grade levels
- Method: train multiple classifiers on manually annotated text
  - Linear regression (Feng et al., 2009)
  - Support vector machines (Petersen & Ostendorf, 2009)
- Features:
  - Lexical features: avg. number of words per sentence, avg. number of syllables per word
  - Syntactic features: parse tree height, noun phrase count, verb phrase count, SBAR count

Discourse features

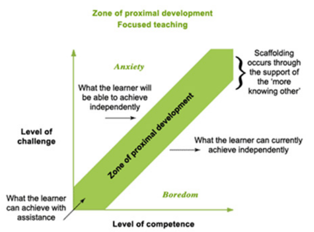
- Discourse features (Pitler & Nenkova, 2008):
  - Vocabulary and discourse relations are the strongest predictors or readability (Wall Street Journal texts)
  - Discourse relations also robustly predict readability rankings (comparisons between two documents)
- Cognitively motivated features for a specific group of users (Feng et al., 2009)
  - Target group: adults with intellectual disabilities
  - Discourse level features: entity density, lexical chains

Document Retrieval for Reading Practice

- Reading proficiency is a widespread problem
  - 29% of high school seniors in public schools across America were below basic achievement in reading in 2005 (Miltsakaki & Troutt, 2008)
  - Low reading proficiency may have dramatic consequences (DuBay, 2004):
    - The strongest risk factor for injury in a traffic accident is the improper use of child safety seats
    - 79 to 94% of car seats are used improperly
    - Installation instructions are too difficult to read for 80% adult readers in the US
- Use readability measures to identify suitable and **authentic** documents, given a reader profile / reading grade

Vygotsky's Zone of Proximal Development

- Materials for **assisted reading** should be harder than the reader's tested reading level, but within the zone of proximal development

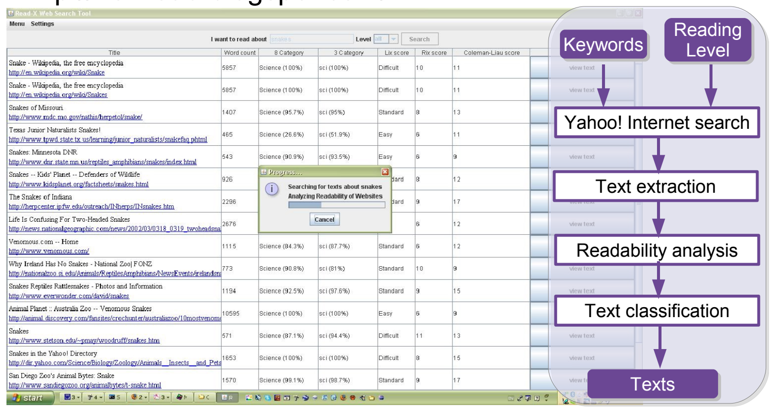


http://www.education.vic.gov.au/

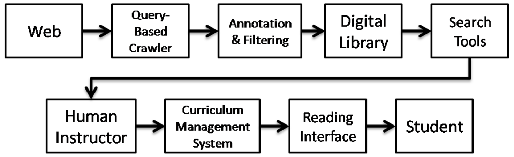
- Materials for **unassisted reading** , e.g. medicine inserts, instructions, should be as easy as possible

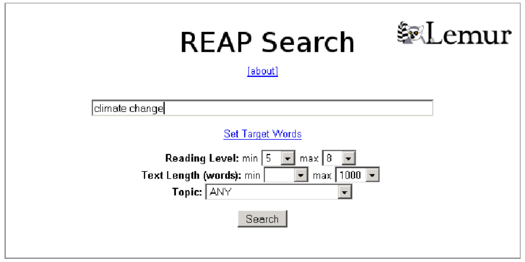
Read-X (Miltsakaki & Troutt, 2008)

- http://net-read.blogspot.com/



REAP search (Heilman et al., 2008)





Text Simplification

- The readability of a text can be improved by transforming it into a simpler text
- Characteristics of manually simplified texts (Petersen & Ostendorf, 2007) :
  - shorter sentences
  - fewer and shorter phrases
  - fewer adjectives, adverbs and coordinating conjunctions
  - nouns are less often replaced with pronouns

    Original text: Congress gave Yosemite the money to repair damage from the 1997 flood.

    Abridged text: Congress gave the money after the 1997 flood

Automatic Text Simplification

- Related techniques: summarisation and sentence compression
- **Syntactic simplification**:
  - Removal or replacement of difficult syntactic structures, using hand-built transformational rules applied to dependency and parse trees (Carroll et al., 1999; Inui et al., 2003)
- **Lexical simplification**:
  - Goal: replace difficult words with simpler ones (Carroll et al., 1999; Lal & Rüger, 2002)
  - Difficult words are identified using the number of syllables and/or frequency counts in a corpus
  - Choose the simplest synonym for difficult words in WordNet

Vocabulary Assistance for Reading

- Overall goal: support vocabulary acquisition during reading for:
  - children, who learn to read (Aist, 2001)
  - foreign language learners, who read texts in a foreign language
- Problem: a word's context may not provide enough information about its meaning
- Solution: augment documents with dynamically generated annotations about (problematic) words

Selection of Target Words

TECHNISCHE UNIVERSITÄT DARMSTADT

- All words are annotated
- Annotate selected words
  - Manually selected target words
  - Automatically selected target words
    - (Aist, 2001):
      - Words with few senses in WordNet (to avoid WSD)
      - Not a trivially easy word: three or more letters long, not in a stop list of function words, not a number
      - Not a proper noun
      - Socially acceptable, e.g. no secondary slang meanings
    - (Mihalcea & Csomai, 2007): keyword extraction methods

Resources for Vocabulary Assistance

TECHNISCHE UNIVERSITÄT DARMSTADT

- **WordNet** (Aist, 2001):
  - Extraction of comparison words for a target word: antonym, hypernym, synonym
  - Generation of factoids:
    - eggshell can be a kind of natural covering
  - Problems:
    - some of the automatically generated factoids are too obscure or do not match the sense of the word used in the original text
    - some of the comparison words may be harder to understand than the target word
    - hypernyms do not always capture the key elements of the meaning of a word

Resources for vocabulary assistance

TECHNISCHE UNIVERSITÄT DARMSTADT

- Collaborative and online resources, e.g. **Wikipedia, Wiktionary, Beolingus, ...**

http://lingro.com/

Wikipedia and Wiktionary as Lexical-Semantic Resources

TECHNISCHE UNIVERSITÄT DARMSTADT

+

This image is licensed under the GFDL. It is based on Bild:Foerderturm-Kamen.jpg.

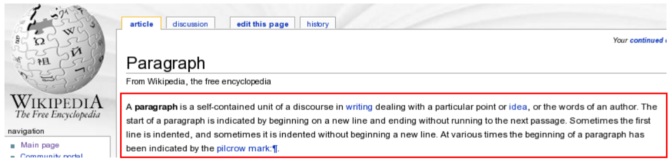
= Lexical semantic resources

- Structure Mining
- Content Mining
- Usage Mining

Wikipedia Article Page
First paragraph

TECHNISCHE UNIVERSITAT DARMSTADT

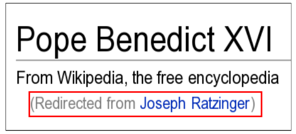
- First paragraph
  - Definition / Gloss



Wikipedia – Redirect Pages

TECHNISCHE UNIVERSITAT DARMSTADT

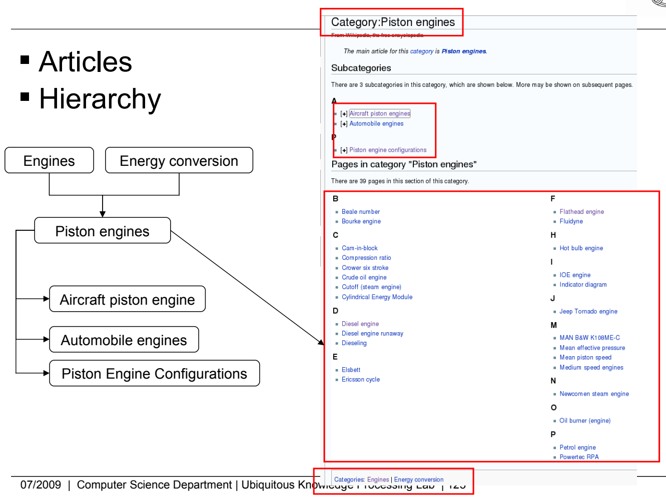
- Synonyms
  - *Pope Benedict XVI*
  - *Joseph Ratzinger*
  - *Joseph Cardinal Ratzinger*
- Spelling variations
  - *Benedict the Sixteenth*
  - *Benedict the 16th*
  - *Benedict 16th*
  - *Benedict 16*
  - *Benedict XVI*
  - *Benedict xvi*
- Misspellings
  - *Josef Ratzinger* (instead of *Joseph)*
- Abbreviations
  - *PB16*



Wikipedia – Categories

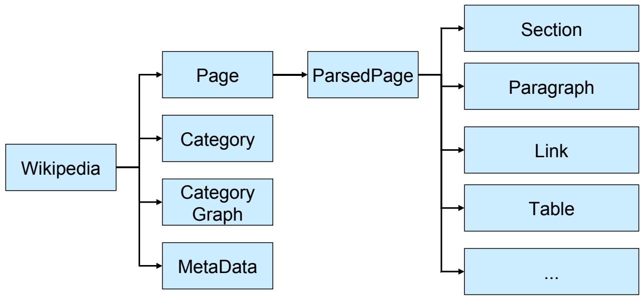
TECHNISCHE UNIVERSITAT DARMSTADT

- Articles
- Hierarchy



JWPL – Wikipedia API

TECHNISCHE UNIVERSITAT DARMSTADT



- Freely available for research purposes
  - http://www.ukp.tu-darmstadt.de/software/

Wiktionary as Lexical-Semantic Resource

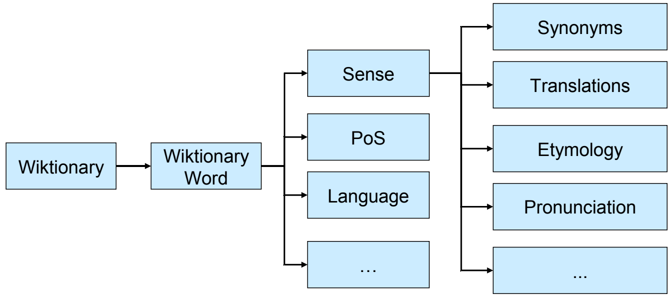
TECHNISCHE UNIVERSITÄT DARMSTADT

- Language
- Etymology
- Pronunciation
- Part-of-speech
- Word senses
- Synonyms
- Derived terms
- Translations

- Abbreviations, Antonyms, Categories, Collocations, Examples, Glosses, Hypernyms, Hyponyms, Morphology, Quotations, Related terms, Troponyms

JWKTL – Wiktionary API

TECHNISCHE UNIVERSITÄT DARMSTADT



- Freely available for research purposes
  - http://www.ukp.tu-darmstadt.de/software/

Wikify! (Mihalcea & Csomai, 2007)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Aim: link keywords (important concepts) in a document to the corresponding Wikipedia page
- Keyword extraction
  - Ranking: tf.idf, χ^2 independence test, keyphraseness
- Word Sense Disambiguation to identify the target Wikipedia page:
  - Lesk algorithm: measure of contextual overlap between the Wikipedia page of the ambiguous word / phrase and the context where the ambiguous word / phrase occurs
  - Machine Learning classifier

Spelling Error Detection and Correction

TECHNISCHE UNIVERSITÄT DARMSTADT

- Aim: identify and correct spelling errors

- Types of spelling errors:
  - **Non-word spelling errors**
    occured instead of occurred
    ater instead of after, later, alter, water, ate
  - **Word conflation or splitting**
    - ofthe, understandhme
    - sp ent, th ebook
  - **Malapropisms**: real-word spelling errors in open-class words
    diary – dairy
    there – their – they're

Research Problems (Kukich, 1992)

TECHNISCHE UNIVERSITAT DARMSTADT

- **Non-word error detection**
  - From the early 1970s to the early 1980s
  - Focus on efficient pattern-matching and string comparison techniques
- **Isolated-word error correction**
  - Started in the early 1960s
- **Context-dependent word correction**
  - Started in the early 1980s
  - Use of statistical language models

Textbook overviews: (Jurafsky & Martin, 2008; Manning, Raghavan and Schütze, 2008)

Non-word Error Detection

TECHNISCHE UNIVERSITAT DARMSTADT

- ***n*-gram analysis**:
  - n-gram = n-letter sub-sequences of words or strings
  - examine each letter n-gram in an input string
  - find the n-gram in a table of n-gram statistics compiled from a corpus of text
  - highly infrequent n-grams indicate probable misspellings
  - especially useful for optical character recognition devices
- **Dictionary lookup**:
  - check if an input string appears in a dictionary of acceptable words
  - techniques: hash tables, tries, finite-state automata, Aho-Corasick algorithm, ternary search trees

Isolated Word Error Correction

TECHNISCHE UNIVERSITAT DARMSTADT

1) Detection of errors in single words, out of context

2) Generation of candidate corrections
   - Distance/Proximity metric between the correct word and the erroneous word
   - Minimum edit distance: minimum number of editing operations (i.e., insertions, deletions, and substitutions) needed to transform one string into another

l	e		v	e	n	s	h	t	e	i	n
o	=	+	o	=	=	=	-	=	=	=	=
m	e	i	l	e	n	s		t	e	i	n

or

l	e	v		e	n	s	h	t	e	i	n
o	=	o	+	=	=	=	-	=	=	=	=
m	e	i	l	e	n	s		t	e	i	n

Distance = 4

"=" Match; "o" Substitution; "+" Insertion; "-" Deletion (c) www.levenshtein.net

3) Ranking of candidate corrections based on the distance/proximity metric or occurrence counts

Isolated Word Error Correction

TECHNISCHE UNIVERSITAT DARMSTADT

Problem: even humans do not achieve 100% accuracy levels, given isolated misspelled strings (Kukich, 1992):

- vver → over, ever, very?
- wekk → week, well, weak?

Context-dependent Error Correction

- Also called context-sensitive spelling correction
- Aim: correct real-word spelling errors, which cannot be identified by dictionary lookup
- Between 25% and 40% of spelling errors are valid English words (Kukich, 1992)
- Use the **context** to help detect and correct spelling errors
- Based on language models

Spelling Correction for Foreign Language Learners (Heift & Rimrott, 2007)

- 80% of the misspellings produced by non-native writers of German are due to insufficient command of the foreign language:
  Metz for Fleisch (from Metzger)
  tanzed for tanzte (from danced)
- These errors are difficult to correct for generic spell checkers → need for rules that are geared towards common L2 errors
- Importance of **feedback**: learners are more likely to correct a mistake if the feedback contains explicit information on the error and correction suggestions

Grammar Checking

- Tasks:
  - **Grammatical error detection**: identify sentences which are grammatically ill-formed
  - **Grammatical error correction**: correct grammatically ill-formed sentences
- Methods:
  - **Rule-based checking**: use of manually written rules
  - **Syntax-based checking**: use the output of a parser
  - **Statistics-based**: use statistical information about n-gram frequencies
  - Many methods focus on a specific part-of-speech, e.g. prepositions

Grammatical Error Types

- According to (Nicholls, 1999, quoted by Chodorow & Leacock, 2000):
  - Insertion of an unnecessary word: *affect **to** their emotions
  - Deletion of a word: *opportunity of job
  - Word or phrase that needs replacing: ***every jobs**
  - Word use in the wrong form: *knowledge**s**
- Grammatical difficulties for ESL learners:
  - Prepositions: *arrive to the town, *most of people, *He is fond this book (Chodorow et al., 2007)
  - Verb forms: I can't *skiing well, I don't want *have a baby (Lee & Seneff, 2008)
  - Articles

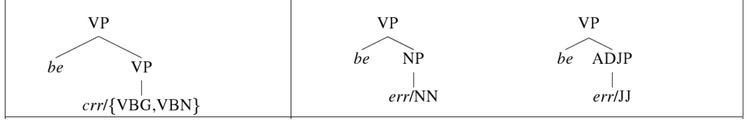
Rule-based Grammar Checking

- Analyse errors in a corpus and write rules to identify and correct these errors, based on POS information
- Rule patterns should not occur in correct sentences
- Examples:
  - Language Tool (Naber, 2003)
    - Open Source language checker
    - Rules are defined in XML configuration files and include feedback messages
  - GRANSKA (Eeg-Olofsson & Knutsson, 2003)
    - Rules expressed in a specific rule language
    - Recall = 25%, Precision = 100%

Syntax-based Grammar Checking

- Template-matching on parse trees (Lee & Seneff, 2008)
  - Automatic introduction of verb form errors in a corpus
  - Parsing of the corpus
  - Identification of templates in the "disturbed" parse trees

Expected Tree $\{\langle usage \rangle, ...\}$	Tree disturbed by substitution $[\langle crr \rangle \rightarrow \langle err \rangle]$
$\{\mathrm{ING}_{prog}, \mathrm{ED}_{pass}\}$	*A dog is [sleeping→sleep]. I'm [living→live] in XXX city.*



Statistics-based Grammar Checking

- Detection of unfrequent sequences of words and/or POS tags:
  - POS **bigrams** (Atwell, 1987)
  - POS tags and function words **n-grams** (Chodorow & Leacock, 2000)

- Machine learning:
  - Maximum entropy model trained with contextual features and combined with rule-based filters (Chodorow et al., 2007)
  - Machine learning model based on automatically labelled sequential patterns (Sun et al., 2007)

Classification based approach

- Method: train a classifier on grammatically correct text to predict which preposition / determiner is correct in a given context (Gamon et al., 2008; De Felice & Pulman, 2008)
- Example contextual features (De Felice & Pulman, 2008):

Head noun	'apple'
Number	singular
Noun type	count
Named entity?	no
WordNet category	food, plant
Prep modification?	yes, 'on'
Object of Prep?	no
Adj modification?	yes, 'juicy'
Adj grade	superlative
POS ± 3	VV, DT, JJS, IN, DT, NN

Table 1: Determiner feature set for *Pick* ***the*** *juiciest apple on the tree.*

POS modified	verb
Lexical item modified	'drive'
WordNet Category	motion
Subcat frame	pp_to
POS of object	noun
Object lexical item	'London'
Named entity?	yes, type = location
POS ± 3	NNP, VBD, NNP
Grammatical relation	iobj

Table 2: Preposition feature set for *John drove* ***to*** *London.*

The Tip of the Tongue Problem

Writers may want to look for words that express a given concept and are appropriate in a given context

Problem: in order to access words in a traditional dictionary, you have to know the word you are looking for

Dictionary Lookup (Ferret & Zock, 2006)

- Tip of the tongue problem:
  - domesticated animal, producing milk suitable for making cheese
  - NOT (cow, buffalo, sheep)
  - → goat
- The *mental* lexicon is a huge network of interconnected words and concepts
- The network is entered through the first word that comes to mind and the target word is retrieved thanks to connecting links

Internal Representation

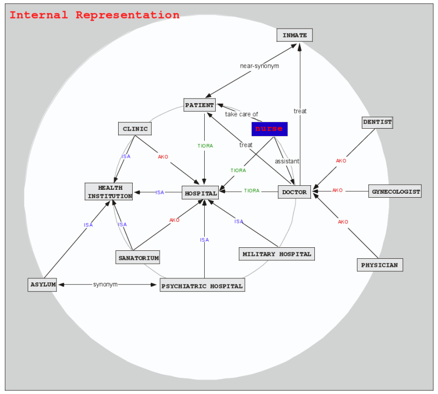
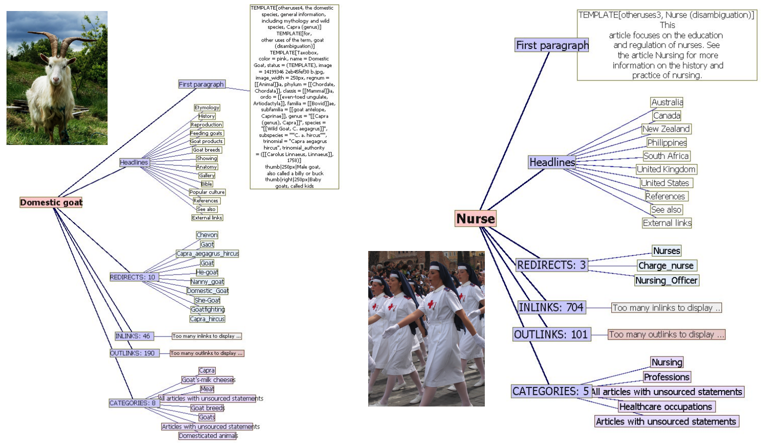


Figure 1: Search based on navigating in a network (internal representation)
AKO: a kind of; **ISA**: subtype; **TIORA**: Typically Involved Object, Relation or Actor.

Wikipedia Graph



Outline

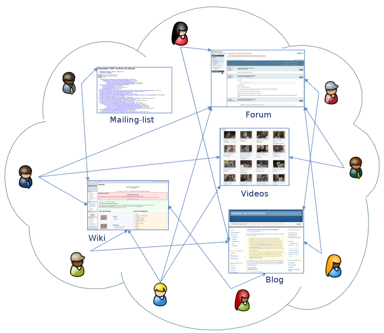
- Introduction: eLearning and NLP
- Automatic generation of exercises
- Assessment of learner generated discourse
- Reading and writing assistance
- **Web 2.0 and computer supported collaborative learning**
- Example e-NLP application: electronic career guidance
- Example e-NLP application: educational QA
- Wrap up and questions

Characteristics of Web 2.0

- Collective intelligence
- Huge amount of data
- Fast growing

- Noise
- Duplicates
- Content of different quality

eLearning 2.0



- **Main characteristics:**
  - Worldwide learning community
  - Educational material produced both by students and teachers
- **Tools:**
  - Wikis
  - Blogs
  - Podcasts
  - Widgets
  - ...

New Learning Paradigms in eLearning 2.0

- Study at any place, any time
  - Several devices may be used for learning: computer, iPod, PDA, etc.
- Authority in educational systems is distributed: collective intelligence and wisdom of the crowds
  - Learn not only from teachers and instructors, but also from peers
- New forms of knowledge organization: tags and folksonomies

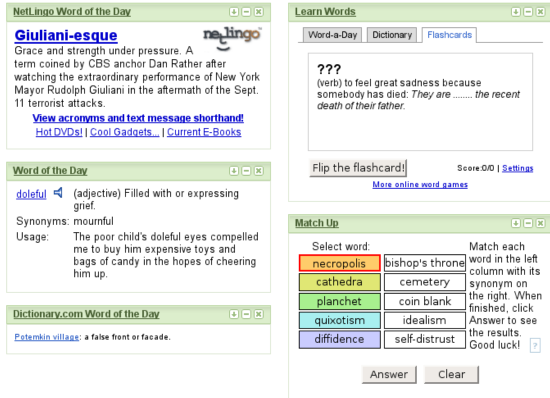
(Bartolomé, 2008)

"CALL 2.0"

TECHNISCHE UNIVERSITAT DARMSTADT

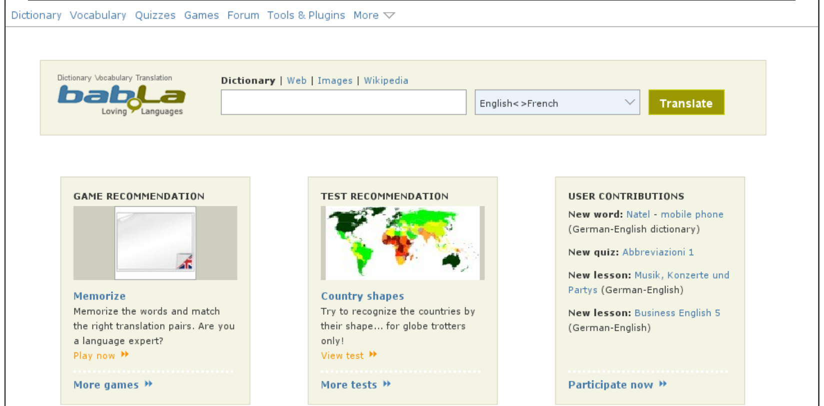
Widgets for CALL

TECHNISCHE UNIVERSITAT DARMSTADT



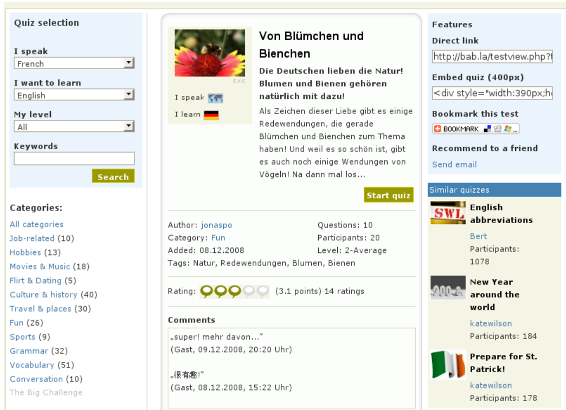
User contributed contents

TECHNISCHE UNIVERSITAT DARMSTADT



User contributed contents

TECHNISCHE UNIVERSITAT DARMSTADT



Use of Web 2.0 Resources

Wikis

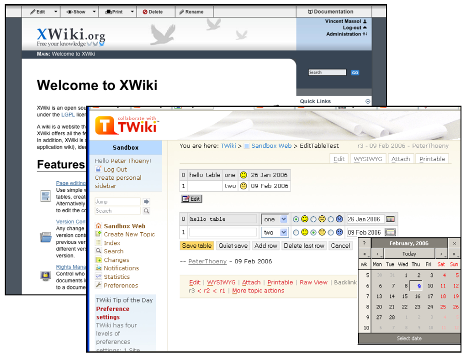
- Goal: build and share knowledge

- Wikis allow users to change contents:
  - collaborative authoring
  - simple wiki markup language
  - stored edit history

- Uses in education:
  - Distribute educational material to students
  - Support student group work
  - Support teacher collaboration

Wiki examples

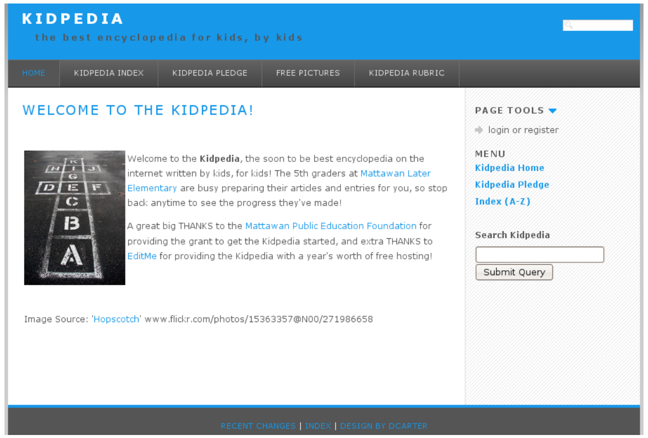
WIKIPEDIA
The Free Encyclopedia



Confluence Enterprise wiki

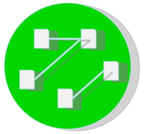
Microsoft Office SharePoint Server 2007

Educational wiki



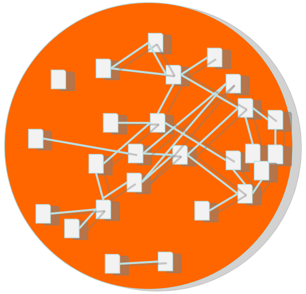
Problems with Wikis

In the beginning …



- Small
- Well **structured**
- Easy to **find** and **add** content

People like it and add **lots** of content



I can't find **anything!**

?

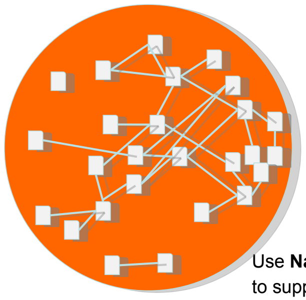
Where do I **put** this?

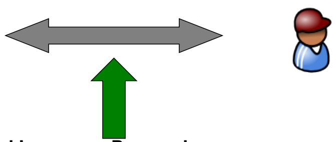
Disorientation and cognitive overload

Wiki User Survey at UKP

- 15 participants
- The two biggest problems
  - Wiki capabilities to re-organize content
  - Finding information
- Confirmed by other studies, e.g.
  - M. Buffa. Intranet Wikis. *Proceedings of the IntraWebs Workshop 2006 at the 15th International World Wide Web Conference WWW 2006.*

UKP's Approach: Wikulu

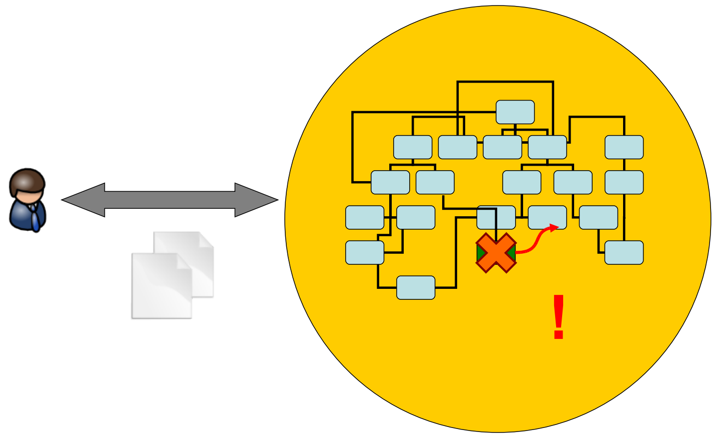




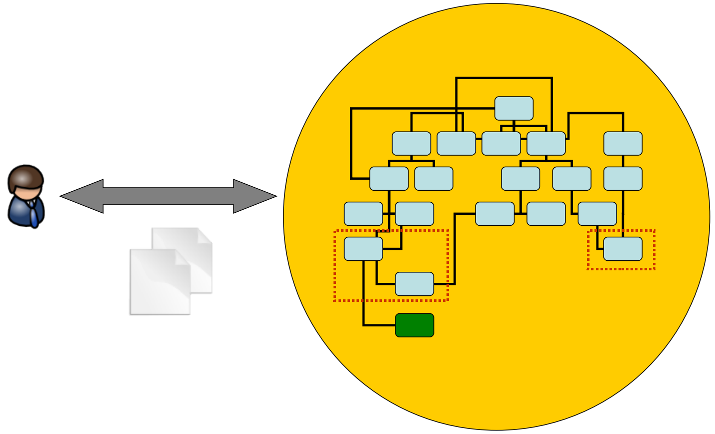
Use **Natural Language Processing**
to support the user by providing suggestions while:
adding, organizing and finding content.

„***Wikulu***“ - Hawaiian for organize [‚*kukulu*'] fast [‚*wiki*']

Adding Content: Detect Duplicate Content



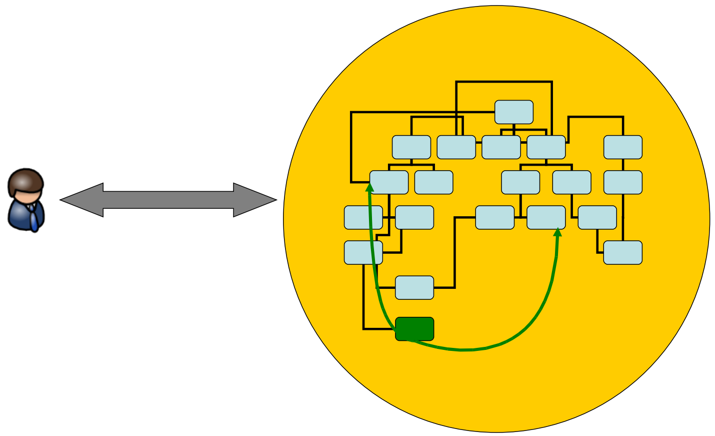
Adding Content: Suggest Points of Insertion



Adding Content: NLP Algorithms

- Text similarity (Gabrilovich & Markovitch, 2007)
  - Highly similar documents might be duplicates
  - ... or possible places for adding the new content
- Text segmentation (Choi et al., 2001)
  - Find specific position for inserting new text by segmenting pages into coherent topics

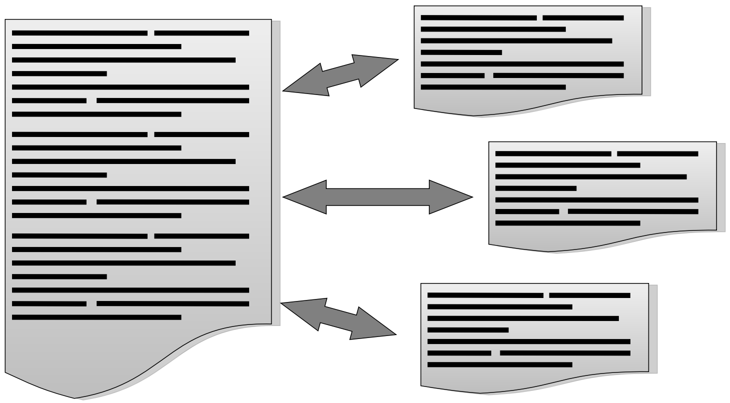
Organizing Content: Suggest Links



Organizing Content: Suggest Tags

CHICAGO, Oct 29 - Kraft Foods Inc and Kellogg Co posted better-than-expected third-quarter profits on Wednesday as price increases and new products helped lift sales in a weak economy. Kraft also stood by its forecasts for 2008 earnings before one-time items as well as for 2009 net income, while Kellogg said its profit this year should hit the high end of its previous targeted range. Both Kraft, the largest North American food maker, and Kellogg, the world's largest cereal company, have taken steps to cut costs and put more money into advertising. Both have also bolstered new product development to attract consumers even as rising commodity costs pushed them to raise prices. Commodities like wheat and energy have become less expensive in recent months, but food companies may not see a big benefit until next year, in part because they lock in their costs months ahead. Kraft, which makes Oreo cookies, Tang breakfast drink and Oscar Mayer hot dogs, reported a profit of 45 cents a share before one-time items, a penny above what analysts polled by Reuters Estimates had expected.

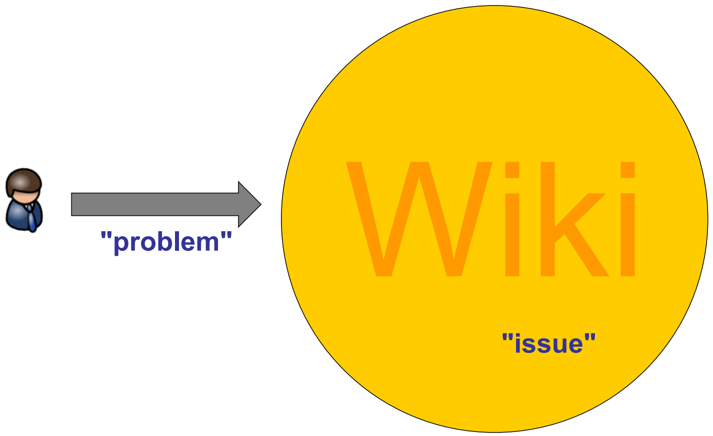
Organizing Content: Suggest Page Split/Merge



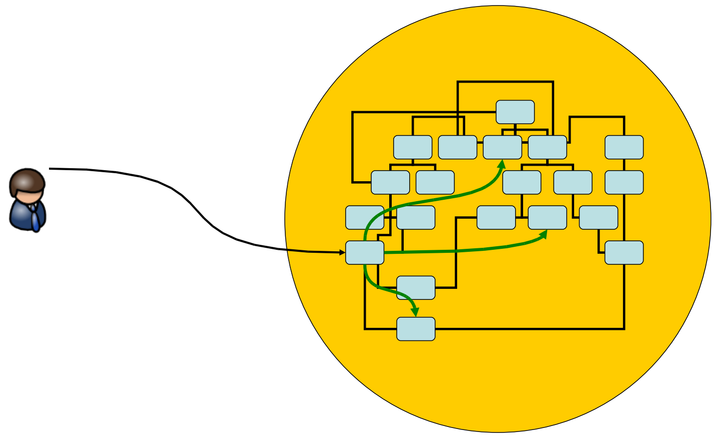
Organizing Content: NLP Algorithms

- Link detection (Green, 1998)
  - Suggest similar content as link target
- Keyphrase extraction (Mihalcea & Tarau, 2004)
  - Propose important keyphrases as possible tags
- Text segmentation
  - Find coherent topics in a page to propose splits
- Text similarity
  - Find scattered pages similar enough to merge

Finding Content: Recall-Oriented Search



Finding Content: Show Related Pages While Browsing



Finding Content: NLP Algorithms

- Text similarity
  - Improve search recall by taking into account term similarity to find additional relevant pages
  - Show related pages while browsing

What is actually the Quality of Web 2.0 Resources?

- Wikipedia:
  - Open edit policy, yet high quality articles (Giles, 2005)
    - 42 entries tested by experts
    - average science entry in Wikipedia contained around four inaccuracies
    - average science entry in Encyclopaedia Britannica contained around three inaccuracies
- Automatic assessment of the quality of these ressources:
  - Social Q&A sites (Jeon et al., 2006; Agichtein et al., 2008)
  - Wikipedia (Druck et al., 2008)
  - Forums (Weimer et al., 2007; Weimer & Gurevych, 2007)

Quality Assessment of User Generated Discourse

- Web 2.0 leads to massive amounts of data
- Users need content of *good* quality
- Current approach
  - Users label the data for quality
  - Labels are used for filtering
- Problems:
  - Happens rarely
  - New item problem
  - Premature negative consensus (Lampe and Resnick, 2004)

Outline

- Introduction: eLearning and NLP
- Automatic generation of exercises
- Assessment of learner generated discourse
- Reading and writing assistance
- Web 2.0 and computer supported collaborative learning
- **Example e-NLP application: electronic career guidance**
- Example e-NLP application: educational QA
- Wrap up and questions

TECHNISCHE UNIVERSITÄT DARMSTADT

The SIR project:

Semantic Information Retrieval for Electronic Career Guidance

Deutsche Forschungsgemeinschaft DFG

funded by the German Research Foundation

Electronic Career Guidance

TECHNISCHE UNIVERSITÄT DARMSTADT

Essay about professional interests

Descriptions of professions

Query

Documents

Information Retrieval

1. ...
2. ..
3. ...

Ranked List of Professions

Problems of Standard Information Retrieval

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Standard search engines**
  1. Return many irrelevant documents (low precision)
  2. Miss many relevant documents (low recall)
- **Why is this the case?**
  - Pure keyword search is often out of context (e.g., *apple, jaguar*)
  - Vocabulary gap:
    - Words are confused with their meaning (*car* = *automobile*)
    - Related words are not considered

Vocabulary Mismatch Problem

TECHNISCHE UNIVERSITÄT DARMSTADT

Semantic Relatedness

Essay

I like baking cakes...

Profession ...

Profession 2

Profession 1

Profession ...

Profession ...

Profession ...

Profession 3

...pastries...
...confectioner...
...food processing industry

Where Does the Information Come From?

TECHNISCHE UNIVERSITÄT DARMSTADT

Knowledge Sources	Wikipedia The Free Encyclopedia	Wiktionary [ˈwɪkʃənrɪ] n., a wiki-based Open Content dictionary	GermaNet
Concepts	Article Titles	Entry Titles	Synsets
Textual Representation	Article Text	Entry Information	Pseudo Glosses

Lexical Semantic Knowledge

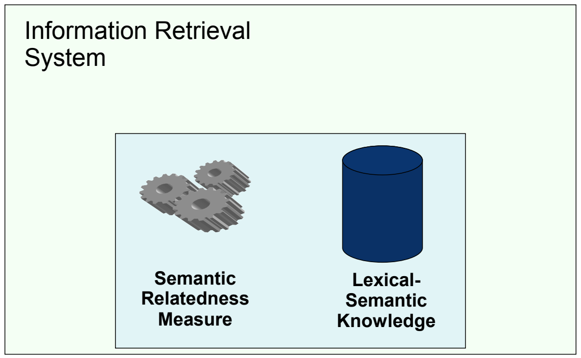
TECHNISCHE UNIVERSITÄT DARMSTADT

- GermaNet: German lexical-semantic wordnet
  - Nouns, verbs, adjectives
  - 27,824 noun synsets, 8,810 verb synsets, 5,141 adjective synsets
  - 60,646 words in synsets
- Wikipedia
  - Free online collaboratively constructed encyclopedia
  - Articles, links, categories (Zesch, Gurevych & Mühlhäuser, 2007)
- Wiktionary
  - Free online collaboratively constructed dictionary
  - Words, categories, semantic relations
- http://www.ukp.tu-darmstadt.de/software/WikipediaAPI

• Semantic relatedness (SR) as measure for document relevance

TECHNISCHE UNIVERSITÄT DARMSTADT

- Semantic relatedness (SR) as measure for document relevance



Semantic Relatedness Measures

TECHNISCHE UNIVERSITÄT DARMSTADT

- Path length (PL)
- Pseudo glosses based (Gurevych, 2005)
- Information content based
  - Resnik (1995)
  - Jiang & Conrath (1997)
  - Lin (1998)
- ESA - Explicit semantic analysis (Gabrilovich & Markovitch, 2007)

ESA: Words are Represented as Concept Vectors

TECHNISCHE UNIVERSITAT DARMSTADT

taxicab

automobile	0.9
drive	0.8
fast	0.6
hire	0.8
New York	0.7
passenger	0.9
SUV	0.1
taxi	1.0
transport	0.9
yellow	0.8

Yellow

In some countries, **taxicabs** are commonly yellow. This practice began in Chicago, where taxi entrepreneur John Hertz painted his taxis yellow based on a University of Chicago study alleging that yellow is the color most easily seen at a distance.

Computing Similarity

TECHNISCHE UNIVERSITAT DARMSTADT

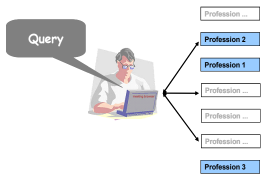
taxicab			truck
automobile	0.9	0.7	automobile
drive	0.8	0.8	drive
fast	0.6	0.2	fast
hire	0.8	0.1	hire
New York	0.7	0.0	New York
passenger	0.9	0.1	passenger
SUV	0.1	0.0	SUV
taxi	1.0	0.0	taxi
transport	0.9	0.9	transport
yellow	0.8	0.1	yellow

$\overline{V}_{taxicab} \times \overline{V}_{truck} =$ Semantic Relatedness

Experiments in Information Retrieval

TECHNISCHE UNIVERSITAT DARMSTADT

"On the other hand, I prefer working with computers, I can program in C, Python and VB and I could therefore imagine working in the software industry."



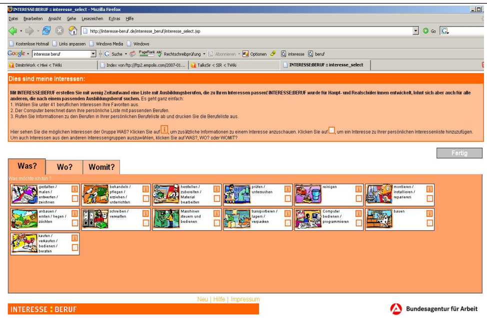
· Topics - 30 essays of human subjects about professional interests
· Queries:
  - Nouns, Verbs, Adjectives
  - Nouns
  - Keywords (set of 41 keywords)

Document Collection

TECHNISCHE UNIVERSITAT DARMSTADT

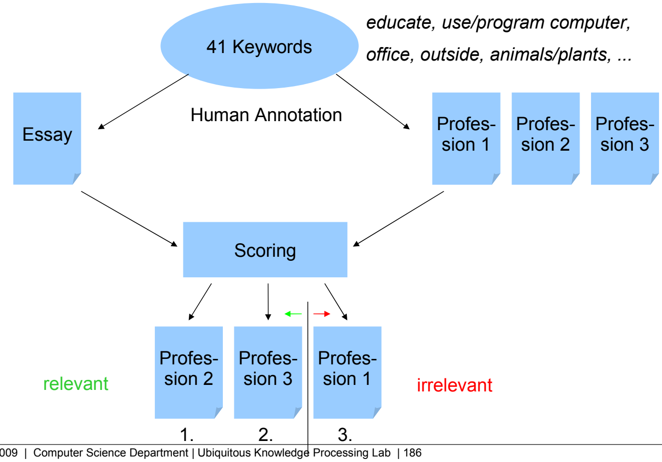
- Provided by the German Federal Labour Office
  - Descriptions of 4,000 professions and 1,800 vocational trainings
  - Prepared by professionals
- Evaluation on 529 descriptions of vocational trainings
- Using parts which describe profession itself, but not training or administrative details

"Gold Standard"



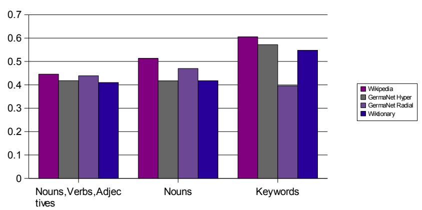
- 41 keywords in 3 categories
- Ranked list of professions for each topic
  - Automatically extracted from knowledge base
  - Used for creating relevance judgments

Relevance Judgments



Results

- Semantic methods lead to up to 40% improvement of search results
- Comparison of the contributions of different ressources
  - Wikipedia scores best



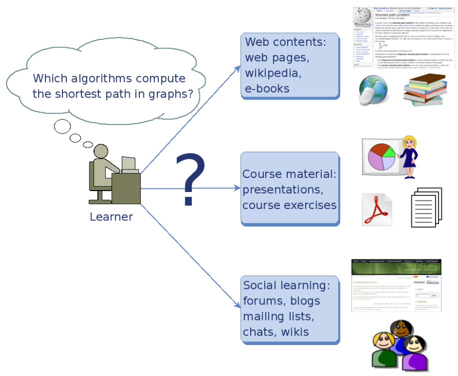
Outline

Introduction: eLearning and NLP
Automatic generation of exercises
Assessment of learner generated discourse
Reading and writing assistance
Web 2.0 and computer supported collaborative learning
Example e-NLP application: electronic career guidance
Example e-NLP application: educational QA
Wrap up and questions

QA-EL
Question Answering for E-Learning

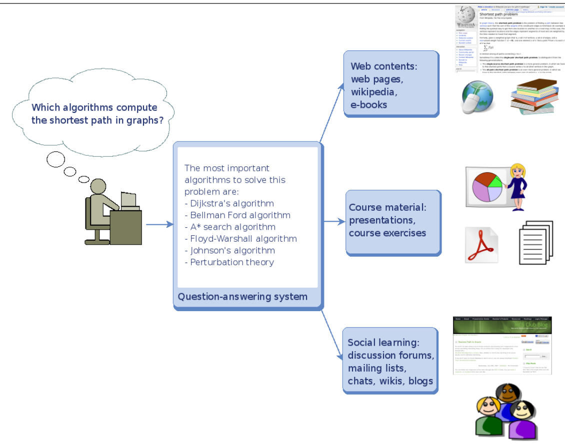
TECHNISCHE UNIVERSITÄT DARMSTADT

Motivation: Information overload in E-Learning



QA-EL
Question Answering for E-Learning

TECHNISCHE UNIVERSITÄT DARMSTADT



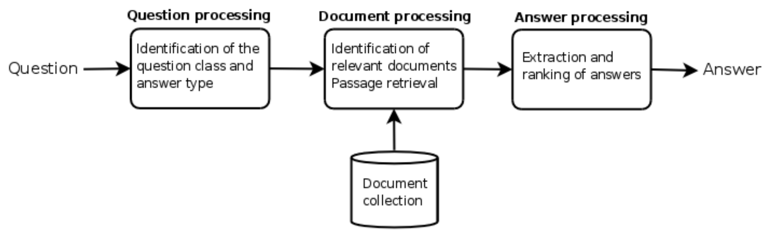
Question Answering (QA) vs. Information Retrieval (IR)

TECHNISCHE UNIVERSITÄT DARMSTADT

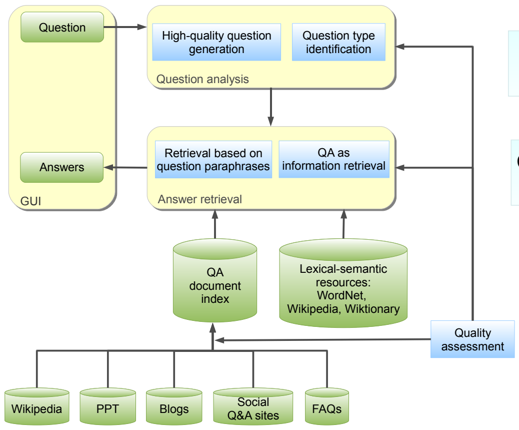
- INPUT:
  - Natural language **questions** and not keyword-based queries:
    - QA: How long do polar bears live?
    - IR: polar bears life span
- OUTPUT:
  - Precise and concise **answers**, not whole documents
    - QA: In the wild, polar bears live an average of 15 to 18 years, although biologists have tagged a few bears in their early 30s. In captivity, they may live until their mid- to late 30s. One zoo bear in London lived to be 41.
    - IR:
      www.gotpetsonline.com/polar-bear/bear-habitat-polar/polar-bear-life-span.html
      www.starbus.com/polarbear/aboutpb.htm
      www.polarbearsinternational.org/faq/

Conventional QA systems

TECHNISCHE UNIVERSITÄT DARMSTADT



Architecture of an Educational QA System (Gurevych et al., 2009)



Assessment of the quality of questions and answers

Combination of answers from heterogeneous documents

Low Quality Questions

TECHNISCHE UNIVERSITAT DARMSTADT

Factor	Example
⇒ Misspelling	**Hou** to cook pasta?
⇒ Internet slang	How **r** plants used **4** medicine?
⇒ Ill-formed syntax	What Alexander Pushkin famous for?
⇒ Keyword search	Drug classification, pharmacodynamics
⇒ Ambiguity	What is the population of **Washington**?

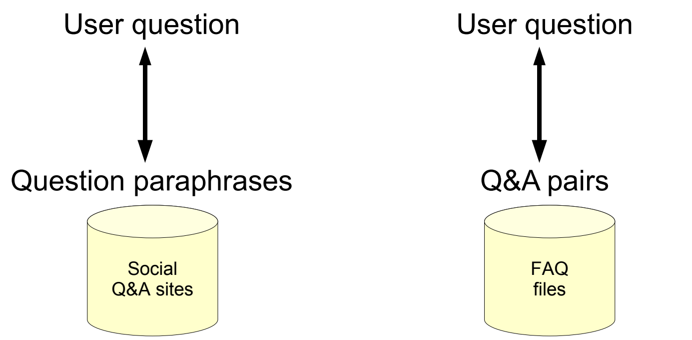
- 755 questions from Yahoo! Answers:
  - 18% misspelled, 8% Internet slang, 20% ill-formed
- Keyword queries are the natural way for most people to look for information
- Ambiguity / Underspecification is harder to identify and is highly context-dependent

K. Ignatova, C. Toprak, D. Bernhard, I. Gurevych. *Generating High Quality Questions from Low Quality Questions*. Workshop on the Question Generation Shared Task and Evaluation Challenge, September 2008.

Question Answering as Reuse

TECHNISCHE UNIVERSITAT DARMSTADT

Question paraphrases | **Information Retrieval**



Question and Answer Repositories

TECHNISCHE UNIVERSITAT DARMSTADT

FAQs

- Questions and answers are compiled and subject to editorial control
- Examples: www.faqs.org

Ask-an-expert Services

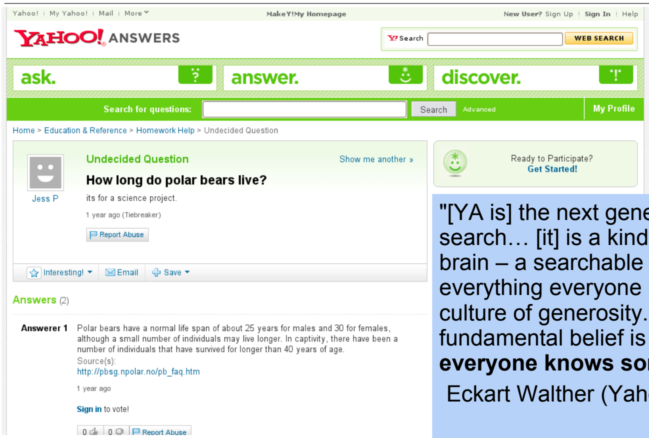
- Provide expert anwers to user questions
- Example: www.madsci.org

Social Q&A sites

- Provide portals where users can ask their own questions and answer questions from other users
- Examples: Yahoo! Answers, WikiAnswers

Example question in Yahoo! Answers

TECHNISCHE UNIVERSITAT DARMSTADT



"[YA is] the next generation of search… [it] is a kind of collective brain – a searchable database of everything everyone knows. It's a culture of generosity. The fundamental belief is that **everyone knows something**"
Eckart Walther (Yahoo research)

WikiAnswers

TECHNISCHE UNIVERSITAT DARMSTADT

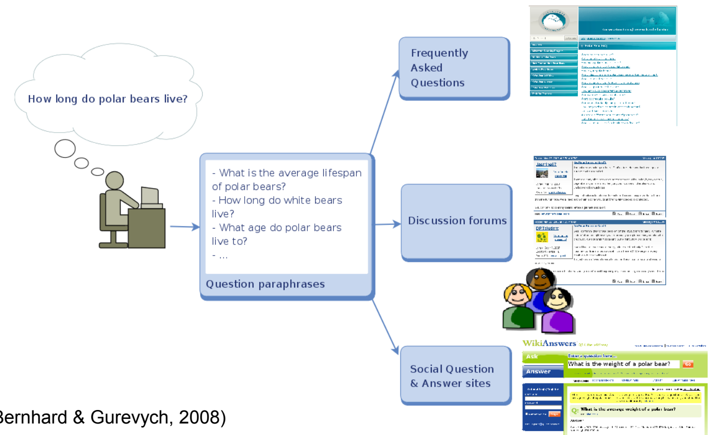
Properties of Social Q&A Sites

TECHNISCHE UNIVERSITAT DARMSTADT

- Managed by the **internet community**, users can:
  - Ask their own questions
  - Answer questions from other systems
- **Ratings** as community mechanism:
  - Points for answers, "Best Answer", oder "thumbs up"
  - Minus points for asking a question
- The American version of *Yahoo! Answers* is the **second-most visited education/reference site on the Internet after Wikipedia** (according to Comscore)

Question Paraphrase Identification

TECHNISCHE UNIVERSITAT DARMSTADT



(Bernhard & Gurevych, 2008)

Question Paraphrase Identification

Input question

Target question

Pre-processing:
- stemming
- lemmatisation
- spelling correction

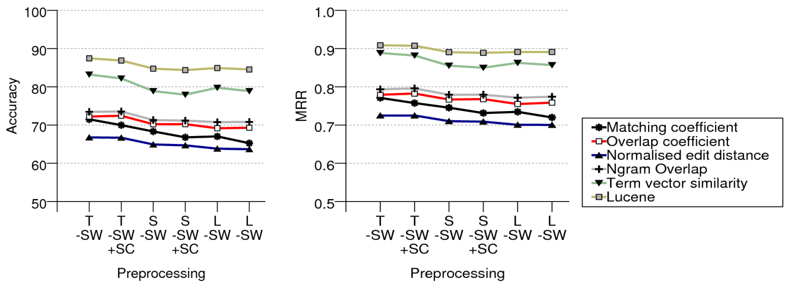
Social Q&A sites

Question similarity measures

String similarity:
- matching coefficient
- overlap coefficient
- edit distance
- ngram overlap

Vector space:
- term vector similarity
- Lucene

Results



- Vector-space based methods outperform string similarity
- Morphological pre-processing and spelling correction do not ameliorate the results

Challenges in Question Paraphrase Identification in Social Q&A Sites

- Spelling errors:
  - How do you become an anes**te**siologist?
  - How many years of medical school do you need to be an anesthesio**lg**ist?
- Vocabulary mismatch:
  - What events occurred in 1919?
  - What important events happened in 1919?
- Solutions:
  - Named entity recognition to identify important tokens in questions
  - Semantic relatedness metrics

Outline

Introduction: eLearning and NLP

Automatic generation of exercises

Assessment of learner generated discourse

Reading and writing assistance

Web 2.0 and computer supported collaborative learning

Example e-NLP application: electronic career guidance

Example e-NLP application: educational QA

Wrap up and questions

NLP has lots to offer

- Resources:
  - Lexical semantic resources, e.g. WordNet
  - Web 2.0 resources, e.g. Wikipedia, Wiktionary
- Tools:
  - Tokeniser and sentence splitting
  - Morphological analysis
  - Part of speech tagging
  - Parsing and chunking
  - Word sense disambiguation
  - Summarisation
  - Keyword extraction

Tasks and applications

- To assist instructors
  - Automatic generation of questions and exercises
  - Assessment of learner-generated discourse
- To assist learners
  - Reading and writing assistance
  - Electronic career guidance
  - Educational question answering
- For all users in the Web 2.0
  - NLP for wikis
  - Quality assessment of user generated contents

What the tutorial has not covered…

- A lot more research is done on:
  - Computer-Assisted Language Learning
  - Intelligent Tutoring Systems
  - Information search for eLearning
  - Educational blogging
  - Annotations and social tagging
  - Analysing collaborative learning processes automatically
  - Learners' corpora and resources
  - eLearning standards, e.g. SCORM

NLP meets educational computing

- Educational applications are challenging for NLP since they place strong quality and robustness requirements on applications
- Interdisciplinary approach:
  - psychology
  - educational computing
  - NLP
  - cognitive and learning sciences
- Emerging types of discourse and learning paradigms in Web 2.0

How to Promote e-NLP?

- Establish an international community
- ACL and AIED associated meeting series
- Related tutorials
- Resources:
  - Bibliography
  - Research groups
  - Projects
  - Annotated corpora
  - Tools

Thank you!

Ubiquitous Knowledge Processing Lab

http://www.ukp.tu-darmstadt.de

References

Automatic Generation of Exercises

— Computer-based Testing and Question Generation —

Duvall, K. Improving Your Test Questions. [Online; visited May 26, 2008]. Center for Teaching Excellence, University of Illinois at Urbana-Champaign. `http://www.oir.uiuc.edu/dme/exams/ITQ.html`.

McKenna, C. and Bull, J. (1999). Designing effective objective test questions: an introductory workshop. [Online; visited May 26, 2008]. CAA Centre, Loughborough University, `http://caacentre.lboro.ac.uk/dldocs/otghdout.pdf`.

— Multiple-choice Questions —

Brown, J. C., Frishkoff, G. A., and Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 819–826, Morristown, NJ, USA. Association for Computational Linguistics.

Heilman, M. and Eskenazi, M. (2007). Application of Automatic Thesaurus Extraction for Computer Generation of Vocabulary Questions. In *Proceedings of Speech and Language Technology in Education (SLaTE2007)*, pages 65–68.

Karamanis, N., Ha, L. A., and Mitkov, R. (2006). Generating Multiple-Choice Test Items from Medical Text: A Pilot Study. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 111–113, Sydney, Australia. Association for Computational Linguistics.

Mitkov, R., Ha, L. A., and Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12(2):177–194.

— Fill-in-the-blank Questions —

Aldabe, I., de Lacalle, M. L., Maritxalar, M., Martinez, E., and Uria, L. (2006). ArikIturri: An Automatic Question Generator Based on Corpora and NLP Techniques. In Ikeda, M., Ashley, K. D., and Chan, T.-W., editors, *Intelligent Tutoring Systems*, volume 4053 of *Lecture Notes in Computer Science*, pages 584–594. Springer.

Coniam, D. (1997). A Preliminary Inquiry Into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests. *CALICO Journal*, 14:15–33.

— Multiple-choice Cloze Questions —

Chen, C.-Y., Liou, H.-C., and Chang, J. S. (2006). FAST: an automatic generation system for grammar tests. In *Proceedings of the COLING/ACL Interactive presentation sessions*, pages 1–4, Morristown, NJ, USA. Association for Computational Linguistics.

Hoshino, A. and Hiroshi, N. (2005). A Real-Time Multiple-Choice Question Generation For Language Testing: A Preliminary Study. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 17–20, Ann Arbor, Michigan. Association for Computational Linguistics.

Lee, J. and Seneff, S. (2007). Automatic Generation of Cloze Items for Prepositions. In *Proceedings of INTERSPEECH 2007*, pages 2173–2176, Antwerp, Belgium.

Liu, C.-L., Wang, C.-H., Gao, Z.-M., and Huang, S.-M. (2005). Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 1–8, Ann Arbor, Michigan. Association for Computational Linguistics.

Smith, S., Sommers, S., and Kilgarriff, A. (2008). Learning words right with the Sketch Engine and WebBootCat: Automatic cloze generation from corpora and the web. In *Proceedings of the Conference of English Teaching and Learning in R.O.C.*

Sumita, E., Sugaya, F., and Yamamoto, S. (2005). Measuring Non-native Speakers' Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 61–68, Ann Arbor, Michigan. Association for Computational Linguistics.

— Matching Test Items —

Brown, J. C., Frishkoff, G. A., and Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 819–826, Morristown, NJ, USA. Association for Computational Linguistics.

— Error Correction Questions —

Chen, C.-Y., Liou, H.-C., and Chang, J. S. (2006). FAST: an automatic generation system for grammar tests. In *Proceedings of the COLING/ACL Interactive presentation sessions*, pages 1–4, Morristown, NJ, USA. Association for Computational Linguistics.

— Item Analysis —

Zurawski, R. M. (1998). Making the Most of Exams: Procedures for Item Analysis. *The National Teaching & Learning FORUM*, 7(6):1–4.

Assessment of Learner-Generated Discourse

— Essay Scoring —

Attali, Y. and Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *Journal of Technology, Learning and Assessment*, 4(3).

Breland, H. M., Jones, R. J., and Jenkins, L. (1994). The College Board vocabulary study. Technical report, College Board Report No. 94–4, New York: College Entrance Examination Board.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., and Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. In *Proceedings of the 17th international conference on Computational linguistics*, pages 206–210, Morristown, NJ, USA. Association for Computational Linguistics.

Burstein, J., Marcu, D., and Knight, K. (2003). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems*, 18(1):32–39.

Burstein, J. and Wolska, M. (2003). Toward evaluation of writing style: finding overly repetitive word use in student essays. In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 35–42, Morristown, NJ, USA. Association for Computational Linguistics.

Elliot, S. M. (2001). IntelliMetric: from here to validity. In *Paper presented at the annual meeting of the American Educational Research Association*, Seattle, WA.

Hearst, M. A. (2000). The Debate on Automated Essay Grading. *IEEE Intelligent Systems*, 15(5):22–37.

Higgins, D., Burstein, J., and Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(2):145–159.

Landauer, T. K., Laham, D., and Foltz, P. (1998). Learning Human-like Knowledge by Singular Value Decomposition: A Progress Report. *Advances in Neural Information Processing Systems*, 10:45–51.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, 47:238–243.

Page, E. B. (1994). Computer Grading of Student Prose, Using Modern Concepts and Software. *Journal of Experimental Education*, 62:127–142.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., and Bhola, D. S. (2002). A Review of Strategies for Validating Computer-Automated Scoring. *Applied Measurement in Education*, 15(4):391–412.

— Plagiarism —

Clough, P. (2000). Plagiarism in Natural and Programming Languages: an Overview of Current Tools and Technologies. Technical report, Internal Report CS-00-05, University of Sheffield.

Clough, P. (2003). Old and new challenges in automatic plagiarism detection. Technical report, National UK Plagiarism Advisory Service.

Martin, B. (1994). Plagiarism: a misplaced emphasis. *Journal of Information Ethics*, 3(2):36–47.

— Short Answer Assessment —

Bailey, S. and Meurers, D. (2008). Diagnosing Meaning Errors in Short Answers to Reading Comprehension Questions. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 107–115, Columbus, Ohio. Association for Computational Linguistics.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In Boitet, C. and Whitelock, P., editors, *Proceedings of ACL/COLING*, San Francisco, California. Morgan Kaufmann Publishers.

Fillmore, C. J. (1976). Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32.

Leacock, C. (2004). Scoring free-responses automatically: A case study of a large-scale assessment. *Examens*, 1(3).

Leacock, C. and Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405.

Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., Kaplan, B., , and Oranje, A. (2005). Online Assessment in Mathematics and Writing: Reports From the NAEP Technology-Based Assessment Project, Research and Development Series. Technical report, National Assessment of Educational Progresss.

Reading and Writing Assistance

— Text Readability —

Brown, J. and Eskenazi, M. (2004). Retrieval of Authentic Documents for Reader-Specific Lexical Practice. In *Proceedings of the InSTIL/ICALL 2004 Symposium on Computer Assisted Learning*, Venice, Italy.

Collins-Thompson, K. and Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

DuBay, W. H. (2004). The Principles of Readability. Costa Mesa, California. Impact Information.

— Document Retrieval for Reading Practice —

Heilman, M., Zhao, L., Pino, J., and Eskenazi, M. (2008). Retrieval of Reading Materials for Vocabulary and Reading Practice. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 80–88, Columbus, Ohio. Association for Computational Linguistics.

Miltsakaki, E. and Troutt, A. (2008). Real Time Web Text Classification and Analysis of Reading Difficulty. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 89–97, Columbus, Ohio. Association for Computational Linguistics.

— Text Simplification —

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999). Simplifying Text for Language-Impaired Readers. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 269–270.

Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T. (2003). Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing*, pages 9–16, Morristown, NJ, USA. Association for Computational Linguistics.

Lal, P. and Rüger, S. (2002). Extract-based Summarization with Simplification. In *Proceedings of the Workshop on Text Summarization at DUC 2002*.

Petersen, S. E. and Ostendorf, M. (2007). Text Simplification for Language Learners: A Corpus Analysis. In *Proceedings of Speech and Language Technology in Education (SLaTE2007)*, pages 69–72.

— Vocabulary Assistance —

Aist, G. (2001). Towards automatic glossarization: automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment. *International Journal of Artificial Intelligence in Education*, 12:212 – 231.

Csomai, A. and Mihalcea, R. (2007). Linking Educational Materials to Encyclopedic Knowledge. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED 2007)*, Los Angeles, CA.

Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, New York, NY, USA. ACM.

Zesch, T., Gurevych, I., and Mühlhäuser, M. (2007). Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. In Rehm, G., Witt, A., and Lemnitzer, L., editors, *Data Structures for Linguistic Resources and Applications*, pages 197–205. Gunter Narr, Tübingen.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of LREC'08*.

— Spell Checking —

Heift, T. and Rimrott, A. (2008). Learner Responses to Corrective Feedback for Spelling Errors in CALL. *System*, 36:196–213.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing*. Prentice Hall. 2nd edition.

Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–439.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

— Grammar Checking —

Atwell, E. S. (1987). How to detect grammatical errors in a text without parsing it. In *Proceedings of the third conference of the European chapter of the Association for Computational Linguistics*, pages 38–45, Morristown, NJ, USA. Association for Computational Linguistics.

Chodorow, M. and Leacock, C. (2000). An Unsupervised Method for Detecting Grammatical Errors. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 140–147.

Chodorow, M., Tetreault, J., and Han, N.-R. (2007). Detection of Grammatical Errors Involving Prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 25–30, Prague, Czech Republic. Association for Computational Linguistics.

Eeg-Olofsson, J. and Knutsson, O. (2003). Automatic grammar checking for second language learners - the use of prepositions. In *Proceedings of NoDaLiDa 2003*.

Felice, R. D. and Pulman, S. G. (2008). A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 169–176, Manchester, UK. Coling 2008 Organizing Committee.

Gamon, M., Gao, J., Brockett, C., Klementiev, A., Dolan, W. B., Belenko, D., and Vanderwende, L. (2008). Using Contextual Speller Techniques and Language Modeling for ESL Error Correction. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, volume 1.

Lee, J. and Seneff, S. (2006). Automatic Grammar Correction for Second-Language Learners. In *Proceedings of INTERSPEECH 2006*, pages 1978–1981.

Lee, J. and Seneff, S. (2008). Correcting Misuse of Verb Forms. In *Proceedings of ACL-HLT-08*, pages 174–182.

Naber, D. (2003). A Rule-Based Style and Grammar Checker. Master's thesis, Technische Fakultät, Universität Bielefeld.

Nicholls, D. (1999). The Cambridge learner corpus - error coding and analysis. In *Summer Workshop on Learner Corpora*, Tokyo, Japan.

Sun, G., Liu, X., Cong, G., Zhou, M., Xiong, Z., Lee, J., and Lin, C.-Y. (2007). Detecting Erroneous Sentences using Automatically Mined Sequential Patterns. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Wagner, J., Foster, J., and van Genabith, J. (2007). A Comparative Evaluation of Deep and Shallow Approaches to the Automatic Detection of Common Grammatical Errors. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 112–121.

— Dictionary Lookup —

Ferret, O. and Zock, M. (2006). Enhancing electronic dictionaries with an index based on associations. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 281–288, Morristown, NJ, USA. Association for Computational Linguistics.

Web 2.0 and Computer Supported Collaborative Learning

— Web 2.0 —

Bartolomé, A. (2008). Web 2.0 and New Learning Paradigms. *eLearning Papers*, 8.

— NLP for Wikis —

Choi, F. Y., Wiemer-Hastings, P., and Moore, J. (2001). Latent Semantic Analysis for Text Segmentation. In Lee, L. and Harman, D., editors, *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 109–117.

Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India.

Green, S. J. (1998). Automated link generation: can we do better than term repetition? *Computer Networks and ISDN Systems*, 30(1-7):75–84.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing Order into Texts. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

— Quality of User-Generated Content —

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding high-quality content in social media. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 183–194, New York, NY, USA. ACM.

Druck, G., Miklau, G., and McCallum, A. (2008). Learning to Predict the Quality of Contributions to Wikipedia. In *Proceedings of the 'Wikipedia and Artificial Intelligence: An Evolving Synergy' Workshop at AAAI-08*.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438:900–901.

Jeon, J., Croft, W. B., Lee, J. H., and Park, S. (2006). A framework to predict the quality of answers with non-textual features. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 228–235, New York, NY, USA. ACM.

Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006). Automatically Assessing Review Helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 423–430, Sydney, Australia. Association for Computational Linguistics.

Weimer, M. and Gurevych, I. (2007). Predicting the Perceived Quality of Web Forum Posts. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 643–648.

Weimer, M., Gurevych, I., and Mühlhäuser, M. (2007). Automatically Assessing the Post Quality in Online Discussions on Software. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume, Proceedings of the Demo and Poster Sessions*, pages 125–128, Prague, Czech Republic. Association for Computational Linguistics.

Electronic Career Guidance

Gurevych, I., Müller, C., and Zesch, T. (2007). What to be? - Electronic Career Guidance Based on Semantic Relatedness. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 1032–1039, Prague, Czech Republic. Association for Computational Linguistics.

Educational Question Answering

Bernhard, D. and Gurevych, I. (2008). Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications, ACL 2008*, pages 44–52, Columbus, Ohio, USA.

Gurevych, I., Bernhard, D., Ignatova, K., and Toprak, C. (2009). Educational question answering based on social media content. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*. (to appear).