

Educational Natural Language Processing Tutorial at COLING'08



Iryna Gurevych, Delphine Bernhard

Educational Natural Language Processing

Presenters



Iryna Gurevych

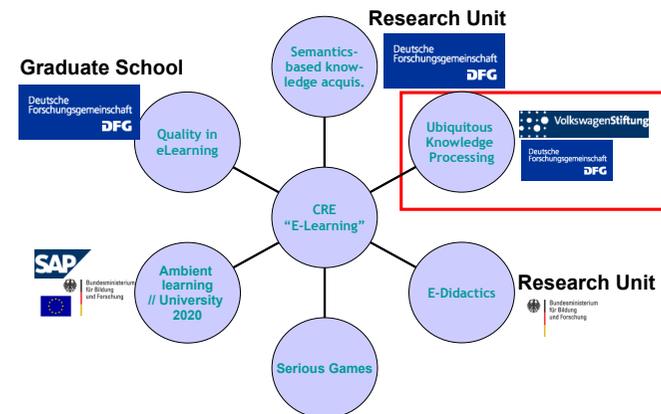


Delphine Bernhard

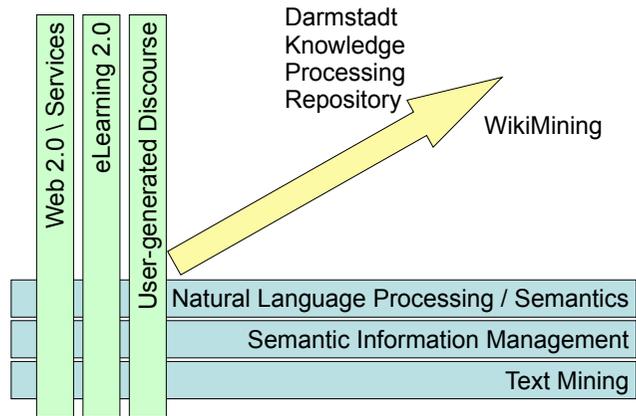
Technische Universität Darmstadt



Center of Research Excellence eLearning 2.0

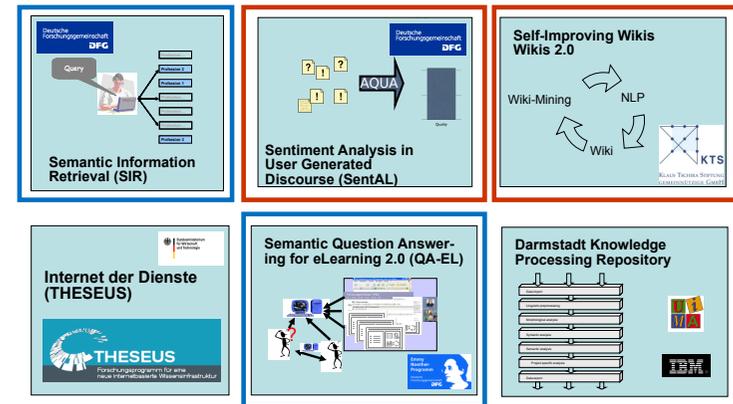


UKP Lab Research Topics



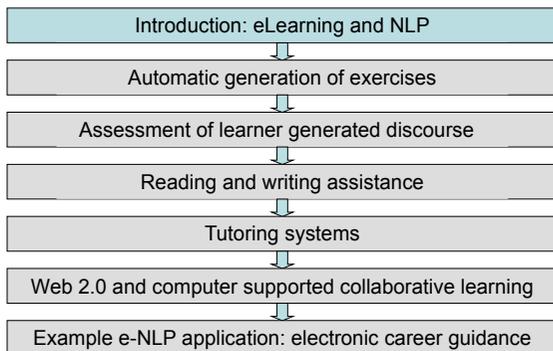
Darmstadt Knowledge Processing Repository
WikiMining

Research Projects and eLearning

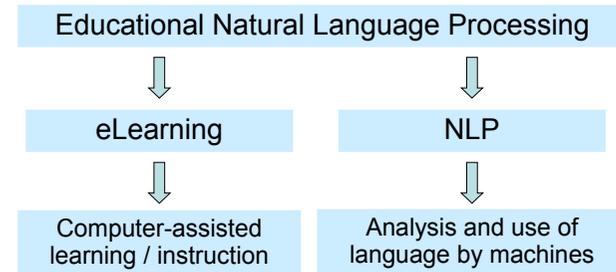


2

Outline



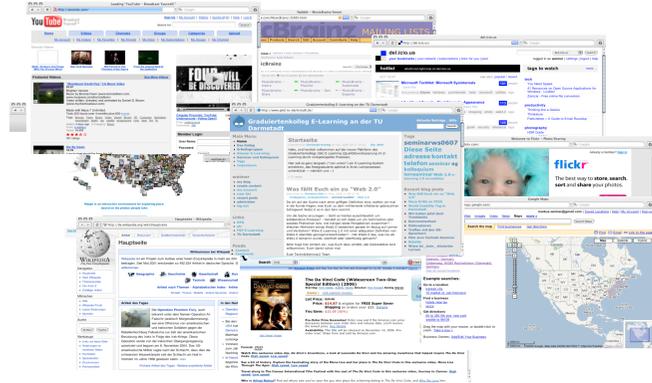
e-NLP



Definition

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

Web 2.0 & eLearning 2.0

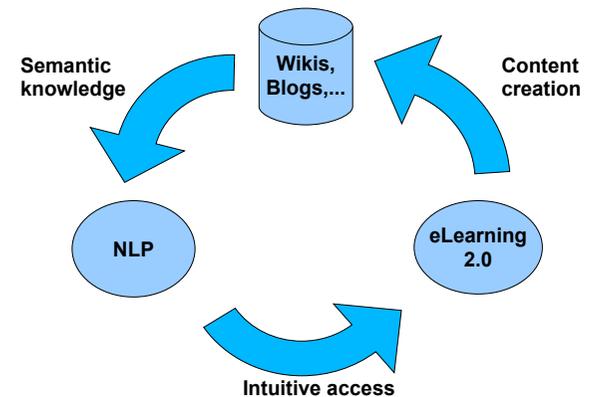


3

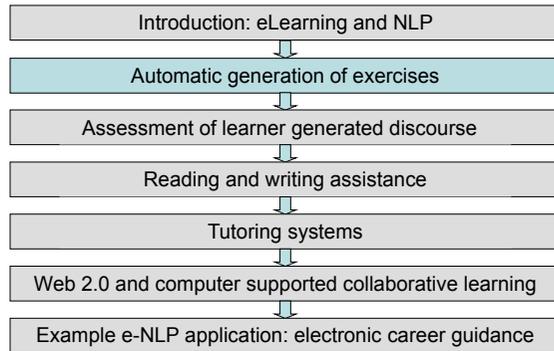
Some Observations

- Creation of large repositories with user generated discourse and user generated metadata
- Using repositories to create structured knowledge bases to improve NLP
- Repositories need advanced information management and NLP to be efficiently accessed

Feedback Loop: NLP & eLearning 2.0



Outline



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.

4

Question Types

- | | |
|--|---|
| <ul style="list-style-type: none">▪ Objective test items<ul style="list-style-type: none">▪ 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:<ul style="list-style-type: none">▪ Fill-in-the-blanks questions▪ Multiple-choice questions▪ Matching questions | <ul style="list-style-type: none">▪ Subjective test items<ul style="list-style-type: none">▪ original answer▪ variable length▪ biased scoring▪ Examples:<ul style="list-style-type: none">▪ Short-answer essays▪ Extended-response essays |
|--|---|

Role of Test Items in Learning

- **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



- **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



- **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

5

Multiple-Choice Questions (MCQ)



- 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
- 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: "verb", Distractors: "modal verbs", "phrasal verbs", "active verbs"

9

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 objective cloze items

Fill-in-the-Blank Examples



- Blank = preposition (Source: <http://www.purl.org/net/WERT/>)

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)
- Using 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	

8

Student Project in the e-NLP Course (Gurevych & Bernhard)



- 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:
- Conclusion: a motivating and interesting project for students

Matching Test Items



- Task: match items on the left column with response items on the right column
- 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

Match Up

Select word: Match each word in the left column with its synonym on the right. When finished, click Answer to see the results. Good luck!

mercurial	arcadian
sanguine	searching
trenchant	ruddy
agile	nimble
bucolic	quicksilver

Match Up Results

Your answers:

mercurial	arcadian
sanguine	searching
trenchant	ruddy
agile	nimble
bucolic	quicksilver

Correct answers:

mercurial	arcadian
sanguine	searching
trenchant	ruddy
agile	nimble
bucolic	quicksilver

mercurial - Quick and changeable in temperament.
Synonyms: [quicksilver](#), [eratic](#), [fickle](#), [volatile](#)
Usage: Her mercurial nature made it difficult to gauge how she would react.

sanguine - Of a healthy reddish color; cheerfully confident.
Synonyms: [robust](#), [ruddy](#), [optimistic](#)
Usage: He had a sanguine complexion that was matched by his cheerful outlook.

trenchant - Having keenness and forcefulness and penetration in thought, expression, or intellect.
Synonyms: [searching](#)
Usage: His trenchant criticism redirected the debate and gave everyone something new to consider.

agile - Characterized by quickness, lightness, and ease of movement; nimble.
Synonyms: [nimble](#), [spry](#), [quick](#)
Usage: She moved quickly and was agile as a gymnast.

bucolic - Of or characteristic of the countryside or its people; rustic.
Synonyms: [rural](#), [arcadian](#), [pastoral](#)
Usage: The illustrations in the book depicted pleasant, bucolic scenes with farmers happily toiling in the fields.

* Correct pairs matched by color, not alignment

Your score is 40% (2 out of 5). Click on any word to learn more.
You may also view the daily [quiz](#) for more Match Up quizzes.
Do you have a website or blog? Add Match Up and other free content with [easy copy and paste code](#).

<http://www.thefreedictionary.com>

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

Wordbank:

verbose infallible obdurate opaque

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

- ___ windows of the jail
- the Catholic Church considers the Pope ___
- ___ and ineffective instructional methods
- the child's misery would move even the most ___ heart

Glosses for specific word senses in WordNet

6

Error Detection Questions

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

Although maple trees are among the most colorful varieties (A)

in the fall, they lose its leaves sooner than oak trees. (B) (C) (D)
- Wrong statements are produced by the distractor generator

Evaluation of Generated Questions

- **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



External assessment

- Evaluate the linguistic and / or factual knowledge of the students before they take the test , e.g. 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 key or not (Heilman & Eskenazi, 2007)
"Do you know the word 'w'?"

Item Analysis



- 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 separately for both groups: p_{upper} and p_{lower}
 - Discrimination index $D = p_{upper} - p_{lower}$

10

Item Analysis



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



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

	Distractor Analysis Data for Upper (U) and Lower (L) Scoring Students			
	Item 1	Item 2	Item 3	Item 4
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.

Well-designed item

Possibly miskeyed

Candidate for removal

Candidate for revision

Efficiency of the Automatic Generation of Test Items



- 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

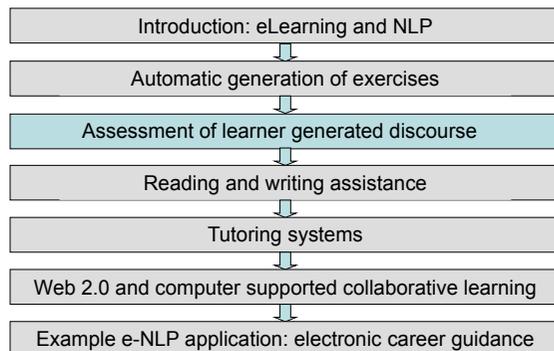
Summary



- 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
 - WSD
 - Term extraction
 - ...

11

Outline



Assessment of Learner Generated Discourse



- Types of learner generated discourse:
 - Emerging in institutional settings, e.g. solutions to exercises
 - Emerging in informal settings, e.g. discussions in forums
- Language forms: written or spoken
- Relevant NLP technologies:
 - Automatic essay grading
 - Detecting meaning errors
 - Plagiarism detection
 - Quality assessment

Importance of Institutional eAssessment



- 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 data to determine the grade and/or making a pass/fail decision

Importance of Free-Text Assessments



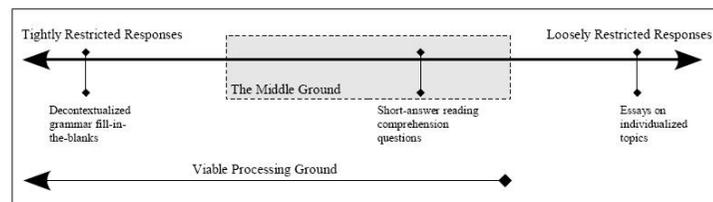
- Advantages over traditional multiple-choice assessments (Bennett & Ward, 1993)
- Major obstacle is the large cost and effort required for scoring
- Automatic systems:
 - Reduce these costs
 - Facilitate extended feedback to students

12

Learning Exercise Spektrum Model (Bailey & Meurers 2008)



- Proposed in the context of language learning, but applicable to different topics
- We will focus on essay grading



What is an Essay?



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

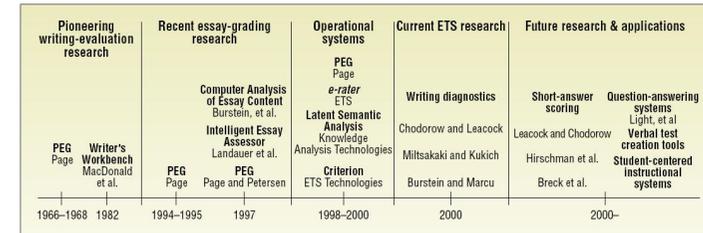
Essay Prompts



- 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.

13

Most Prominent Systems



- **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?



- 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?



- 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)



- 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 objectives:
 - **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
 - **Construct** – 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



- 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)

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



- 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
- Train a system on a large corpus of human annotated essays to identify "good" sequences
 - Mandatory parts, > 3 main ideas, ...

Essay Annotated with Discourse Elements



<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



- Related to word-specific characteristics
 - 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



- 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



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

Future Directions



- 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)

Plagiarism



“Plagiarism is representing the words or ideas of someone else as your own. Examples include, but are not limited to, failing to properly cite direct quotes and failing to give credit for someone else's ideas”.

University of Miami Honor Council, Honor Code

“Plagiarize: To practice plagiarism upon; to take and use as one's own the thoughts, writings, or inventions of another. (With the thing, rarely the person, as object.)”

Oxford English Dictionary Online

How to Avoid it?



- 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 repeat 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

Main Ways of Plagiarism



- "Copy" work:
 - From another student (intra-corporal)
 - From a source outside the corpus of submissions (extra-corporal)
 - Self-plagiarism
- The Internet makes it easier than ever:
 - Download a term paper
 - Fail to give proper credit to the source of an idea
 - Copy extensive passages without attribution
 - Inserting someone else's phrases or sentences (minimally paraphrased) into your own prose and forget to supply a set of quotation marks

Types of Techniques Used to Conceal Copying



- 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 plagiarized 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>

Forms of Plagiarism



- (1) **Word-for-word plagiarism:** direct copying of phrases or passages from a published text without quotation or acknowledgement.
- (2) **Paraphrasing plagiarism:** when words or syntax are changed (rewritten), but the source text can still be recognised.
- (3) **Plagiarism of secondary sources:** when original sources are referenced or quoted, but obtained from a secondary source text without looking up the original.
- (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 authorship:** the direct case of putting your own name to someone else's work

Based on Martin (1994) and Clough (2003)

Typical Plagiarism Indicators



- 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 or quality
- Incoherent text where the flow is not consistent or smooth, which may signal that a passage has been cut-and-pasted from an existing electronic source
- A large degree of similarity between the content of two or more submitted texts. This may include similarity of style as well as content
- Shared spelling mistakes or errors between texts
- Dangling references, e.g. a reference appears in the text, but not in the bibliography
- Use of inconsistent referencing in the bibliography suggesting cut-and-paste

Based on Clough (2003)

String Matching Algorithms



- Most popular plagiarism detection scheme:
 - Finding the overlap of matching subsequences and substrings (consecutive tokens) of length $\geq n$ (where n is derived empirically)
 - The longer n becomes, the more unlikely it is that the same sequence of n tokens (words or characters) will appear in the same order in independently written texts
 - A similarity function is used to capture the degree of overlap between the two texts represented by the sets of n -grams and a chosen threshold above which texts are deemed plagiarised
 - Problem: larger N -grams are rare, difficult to define thresholds

Uniqueness of N-grams (from Clough 2003)

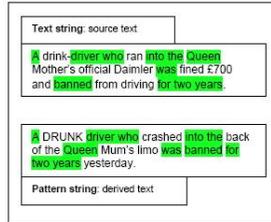


- Figures taken from 769 texts in the METER corpus:

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 1:1 mapping between the tokens in a text pair in such a way that as much of one text as possible is covered with maximal non-overlapping substrings (called *tiles*) from the other.
- This algorithm computes the longest common substrings (greater than length n) between two texts without having to define an n -gram size *a priori*.
- Figure 1 represents a tiling of two sentences after running GST (tiles are highlighted) with a minimum match length of 1 word.

Longest Common Substrings Computed between Two Sentences



- The output of GST algorithm is a set of maximal matches between the text pair: [for two years], [driver who], [into the], [a], [queen], [was] and [banned].
- Different quantitative measures to detect plagiarism, e.g.:
 - the minimum and maximum tile length
 - the average tile length
 - the dispersion of tile lengths
 - a similarity score based on tile length (similar to that for n -gram containment).
- The challenge is to capture these tiling patterns such that derived and non-derived texts are distinguishable.

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

Machine Learning in Plagiarism Detection



- Combining evidence from various sources, e.g.
 - use a Naïve Bayes probabilistic classifier to combine evidence from several measures of similarity taken from a GST tiling and make a decision: derived or not-derived
- *Supervised learning*: training data required (texts which have already been classified as plagiarised or not)
- *Unsupervised learning*: can also be helpful in grouping together texts which exhibit similar characteristics (e.g. *clustering*)

Relaxing the Approach



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

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

NLP in Plagiarism Detection



- 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

Online Internet Plagiarism Services



- Plagiarism.org www.plagiarism.org
 - The largest online plagiarism service available
- IntegriGuard www.integrigaurd.com
- 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

Assessing Short Textual Answers



- Automatic scoring
 - Essays (e-Rater, Burstein and Chodorow, 1999)
 - Longer texts (AutoTutor, Wiemer-Hastings et al., 1999)
- Automatic diagnosis, i.e. content assessment (CAM) on learner data
 - Language learning (Bailey and Meurers, 2008)
 - Error detection in C-rater (Leacock, 2004)
 - 85% accuracy

C-Rater (Chodorow 2004)



- 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



- 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:
 - predicate argument structure
 - resolving the referent of any pronouns in the response
 - regularizing over morphological variation
 - matching on synonyms or similar words
 - resolving the spelling of unrecognized words
- Mapping canonical representations to those of the gold standard responses
 - Rule-based
- 11th grade reading comprehension items
 - Exact agreement with human scorers 84%**

Detecting Meaning Errors (Bailey and Meurers, 2008)



- 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.*
- Positive 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



- Input:
 - Learner's response, one+ target responses, question, source reading passage
 - String-based analysis filter
 - 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)

Technology of CAM



- 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))
 - Chunk level
 - Relation level
- Diagnosis analyzes if the learner's response contains content errors
- 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

17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 85/206

Automatically Scoring Speech



- Non-native speech scoring (Bernstein 1999; Zechner and Bejar, 2006, Zechner et al., 2007)
- SET-10 (Bernstein 1999) focuses on the lower entropy language aspects
 - Tasks such as „reading“ or „repetition“
 - Highly predictable word sequences
- TOEFL Practice Online Speaking test (Zechner et al., 2007)
 - Focus on spontaneous, high-entropy responses
- Test with Heterogeneous Tasks (THT) (Zechner and Xi, 2008)
 - Ranges from reading speech to opinion giving
 - Assess communicative competence

17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 86/206

Test with Heterogeneous Tasks



- Dimensions of assesement:
 - Comprehensibility, accuracy, clarity, coherence, appropriateness
- Evident through:
 - Speaker's pronunciation, fluency, use of grammar and vocabulary, development of ideas, sensitivity to communicative context

17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 87/206

THT Task Types



1. Reading aloud
2. **Picture description (medium-entropy)**
 - Describe a picture in detail
 - Rated on the combined impact of delivery, use of structures, vocabulary, content relevance and fullness (3-point scale)
3. Open-end short-answer questions
4. Constrained short-answer questions
5. Respond to a voice mail
6. **Opinion task (high-entropy)**
 - State an opinion on an issue and support its with reasons, examples, arguments, etc.
 - Rated on the combined impact of fluency, pronunciation, intonation and stress, grammar, vocabulary, content relevance, and cohesion and ides progression (5-point scale)

17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 88/206

Technology of SpeechRater



- Adapt a non-native English speech recognizer (trained on TOEFL Practice Online data) to transcribed THT task responses
- Compute a set of relevant speech features based on the recognition output
- Build a scoring model using a subset of features to predict human scores

Feature Number	Feature Name	Feature Class	Description	Used in
1	hmmscore	Pronunciation	Acoustic Model score: sum of the log probabilities of every frame, normalized for length	Opinion & Picture
2	typesper-second	Fluency & Vocabulary diversity	Number of unique words in response ("types") divided by length of response	Opinion & Picture
3	silences-persecond	Fluency	Number of silences per second	Opinion & Picture
4	repetitions	Fluency	Number of repetitions divided by number of words	Opinion
5	relevance-cos5	Vocabulary & Content	Cosine word vector product between a response and all responses in the training set that have the highest score (5 for the Opinion task)	Opinion
6	relevance-cos3	Vocabulary & Content	Cosine word vector product between a response and all responses in the training set that have the highest score (3 for the Picture task)	Picture

17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 89/206

Evaluation



- Human agreement (kappa): around 0.50 (Picture) and 0.72 (Opinion)
- Opinion task – multiple regression employing Equal, Expert, or Optimal Weights; picture task – CART 5.0 (classification trees)

Model	Multiple Regression (Equal Weights)	Multiple Regression (Expert Weights)	Multiple Regression (Optimal Weights)
Weighted κ	0.53	0.62	0.61
Pearson r Correlation (unrounded)	0.62	0.68	0.69
Pearson r Correlation (rounded)	0.56	0.63	0.63

Table 3. Performance of different weighting schemes on THT scoring model evaluation set for Opinion tasks (generic model)

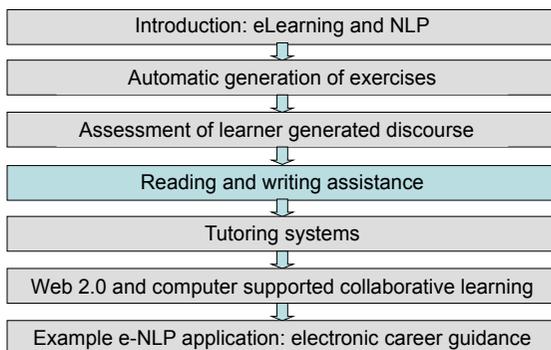
	Generic	Task-specific	Inter-human agreement
Weighted κ	0.51	0.50	0.49
Pearson r Correlation	0.52	0.50	0.50

Table 4. Performance of CART models on THT scoring model evaluation set for Picture tasks (generic model vs. task-specific model)

17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 90/206

23

Outline



17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 91/206

Readability



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

17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 92/206

Traditional Readability Measures

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	- 5 = comic books - 10 = newspapers
SMOG grading	1969	- # words with more than 3 syllables	- 0 to 6 = low-literate - 19+ = post-graduate

Readability Statistics

- 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:
5486 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: 42,7/100
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 Models for Reading Difficulty

- Based on statistical models representing norms, specific populations and individuals (Brown & Eskenazi, 2004)
- Different models are created for each level of reading difficulty
- Features:
 - Lexical features: word unigrams (Collins-Thompson & Callan, 2005; Heilman et al., 2008)
 - Grammatical features: frequency of specific grammatical constructions (Heilman et al., 2007)

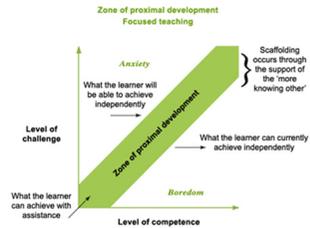
Document Retrieval for Reading Practice

- Reading proficiency is a widespread problem
 - Only 29% of high school seniors in public schools across the USA were proficient in reading according to a 2005 NCES study (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

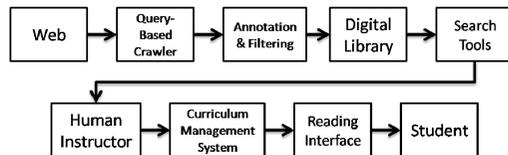
Read-X (Mitsakaki & Troutt, 2008)



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

Title	Word count	Category	Level	Lex score	Column/Lex score
Shaker - Wikipedia, the free encyclopedia	9557	Science (100%)	01 (100%)	Difficult	0.0 11
Shake - Wikipedia, the free encyclopedia	9557	Science (100%)	01 (100%)	Difficult	0.0 11
Shaker of Mirrors	1407	Science (85.7%)	01 (85.7%)	Standard	8 13
Texas Junior-Higher State	465	Science (26.6%)	01 (26.6%)	Easy	9 11
Shaker - Kato Phant - Defender of Wabale	926	Science (80.8%)	01 (80.8%)	Hard	8 12
The Shaker of Indus	2296	Science (82.5%)	01 (82.5%)	Standard	8 15
Life In Confusing Fur Two-Headed Snake	3876	Science (81.1%)	01 (81.1%)	Hard	9 12
Yosemite.com - Home	1115	Science (84.3%)	01 (84.3%)	Standard	9 12
Shaker - Kato Phant - Defender of Wabale	773	Science (80.8%)	01 (80.8%)	Standard	10 8
Shaker - Kato Phant - Defender of Wabale	1194	Science (82.5%)	01 (82.5%)	Standard	8 15
Shaker - Kato Phant - Defender of Wabale	10595	Science (100%)	01 (100%)	Easy	6 8
Shaker - Kato Phant - Defender of Wabale	571	Science (87.1%)	01 (87.1%)	Difficult	11 13
Shaker in the Yahoo! Directory	1653	Science (100%)	01 (100%)	Difficult	8 16
Shaker - Kato Phant - Defender of Wabale	1570	Science (81.1%)	01 (81.1%)	Standard	8 17

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 & Rürger, 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
- Aim: augment documents with dynamically generated annotations about (problematic) words

Selection of Target Words



- 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



- **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



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

<http://lingro.com/>

Die Zuverdienerin

VON WOLFGANG UCHATIUS | © ZEIT online 9.6.2008 - 14:32 Uhr
 SCHLAGWÖRTER: Gleichberechtigung Familie und Partnerschaft
 Gesellschaft

Frauen werden in Deutschland im Durchschnitt schlechter bezahlt als Männer. Das liegt nicht am Frauenhass der Chefs, sondern an alten Rollenbildern

Monatsende in einer deutschen Firma. Die Angestellten kriegen ihre Gehaltsabrechnungen. Die Sachbearbeiterin Frau Müller öffnet ihr Schreiben, der Sachbearbeiter Herr Maier auch. Man kann davon ausgehen, dass Maier ein Mann ist.

Frauen werden in Deutschland im deutlich schlechter bezahlt als Männer, mal auf knapp 30 Prozent wie vom Statistischen Bundesamt, mal irgendwo dazwischen wie von der OECD. In einem aber sind sich alle Studien einig: In kaum einem anderen Industrieland ist der Abstand so groß wie in Deutschland, nirgendwo ist er so dauerhaft, in den vergangenen dreißig Jahren hat er sich kaum verringert.

Sachbearbeiter

1. person responsible (for)
2. advisor
3. consultant
4. official in charge
5. referee

Source: Beolingus

Add to wordlist

Wikipedia and Wiktionary as Lexical-Semantic Resources



WIKIPEDIA
The Free Encyclopedia

+



a multilingual tree encyclopedia

Wiktionary
[ˌwɪkʃənəri] n., a wiki-based Open Content dictionary
WIKIEN [wɪki ˈkɛːrɪ]

=

Lexical semantic resources

- Structure Mining
- Content Mining
- Usage Mining

27

Wikipedia Article Page First paragraph



- First paragraph
- Definition / Gloss

Paragraph
From Wikipedia, the free encyclopedia

A **paragraph** is a self-contained unit of a discourse in writing dealing with a particular point or idea, or the words of an author. The start of a paragraph is indicated by beginning on a new line and ending without running to the next passage. Sometimes the first line is indented, and sometimes it is indented without beginning a new line. At various times the beginning of a paragraph has been indicated by the pīcrow mark ¶.

Wikipedia – Redirect Pages

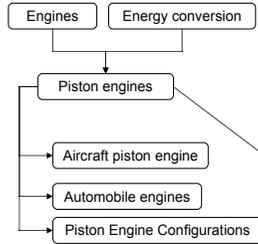


- Synonyms
 - Pope Benedict XVI
 - Joseph Ratzinger
 - Joseph Cardinal Ratzinger
- Spelling variations
 - Benedict the Sixteenth
 - Benedict the 16th
 - Benedict 16th
 - Benedict 16
 - Benedict XVI
 - Benedict xvi
- Misspellings
 - Josef Ratzinger (instead of Joseph)
- Abbreviations
 - PB16

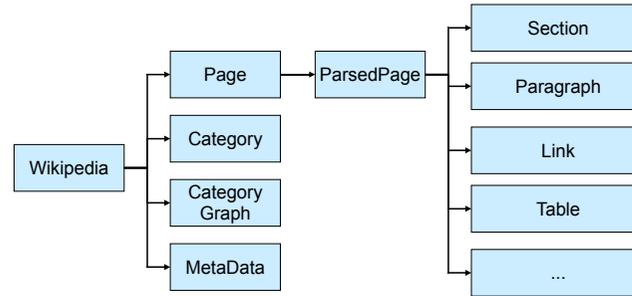
Pope Benedict XVI
 From Wikipedia, the free encyclopedia
 (Redirected from Joseph Ratzinger)

Wikipedia – Categories

- Articles
- Hierarchy



JWPL – Wikipedia API



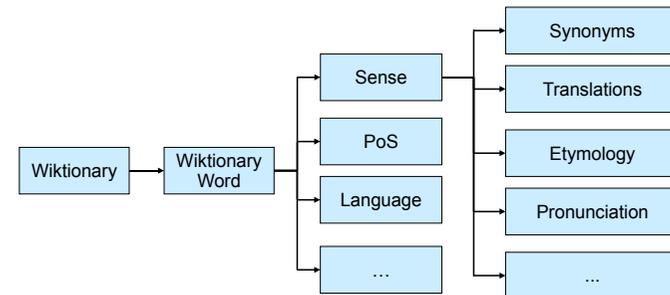
- Freely available for research purposes
- <http://www.ukp.tu-darmstadt.de/software/>

28

Wiktionary as Lexical-Semantic Resource

- 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



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

Wikify! (Mihalcea & Csomai, 2007)



- Aim: link keywords (important concepts) in a document to the corresponding Wikipedia page
- Keywords 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



- Aim: identify and correct spelling errors
- Types of spelling errors:
 - **Non-word spelling errors**
 - ocured 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)



- **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



- **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



- 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 l e v e n s h t e i n
o = + o = = = - = = = or o = o + = = = - = = = Distance = 4
m e i l e n s t e i n m e i l e n s t e i n

"=" 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



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
 - The methods usually focus on a specific part-of-speech

Grammatical Error Types



- According to (Nicholls, 1999):
 - 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: **knowledges*
- 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

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 $\{\langle usage \rangle, \dots\}$	Tree disturbed by substitution $\{\langle crr \rangle \rightarrow \langle err \rangle\}$
$\{ING_{prog}, ED_{pass}\}$	<i>A dog is [sleeping→sleep]. I'm [living→live] in XXX city.</i>
<pre> graph TD VP1[VP] --- be1[be] VP1 --- VP2[VP] VP2 --- crr[crr] VP2 --- VBN["{VBG, VBN}"] </pre>	<pre> graph TD VP3[VP] --- be3[be] VP3 --- NP[NP] NP --- err3[err] NP --- NN[NN] VP4[VP] --- be4[be] VP4 --- ADJP[ADJP] ADJP --- err4[err] ADJP --- JJ[JJ] </pre>

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 rule-based filters (Chodorow et al., 2007)
 - Machine learning model based on automatically labelled sequential patterns (Sun et al., 2007)

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

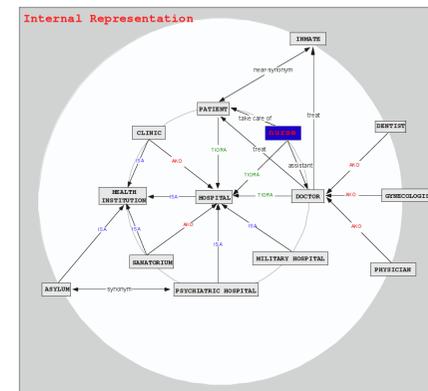
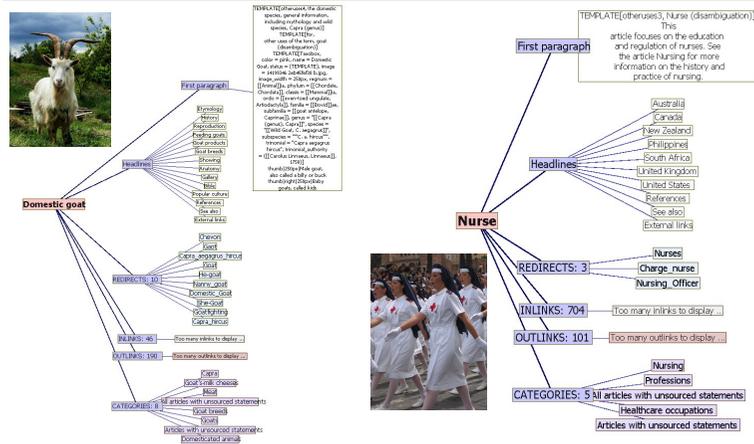
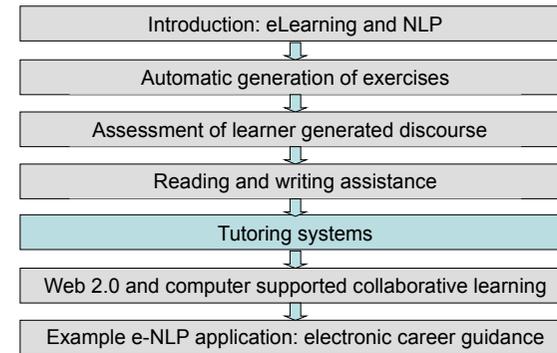


Figure 1: Search based on navigating in a network (internal representation)
 AKO: a kind of, ISA: sub-type, TIORA: Typically Involved Object, Relation or Actor.

Wikipedia Graph



Outline



Intelligent Tutoring Systems with Conversational Dialogue

- Developed during last 25 years, typically the domains of e.g. mathematics, science and technology
- Goal: the ability to engage learners in rich natural language dialogue
- Significant learning gains beyond classroom environments:
 - Learning gains from computer tutors by approximately .3 to 1.0 grade unit (Corbett et al. 1999)
 - Learning gains from human tutors by .4 to 2.3 grade units, though
 - modest domain knowledge
 - no training in pedagogy
 - rare use of sophisticated tutoring strategies

Interaction with Intelligent Tutoring Systems

- System presents problems and questions to learners
- Learner types in / utters answers in natural language
- **Lengthy multi-turn** dialogues as complete solutions / answers evolve

Research on ITS



- CIRCSIM (Evens and Michael 2006)
 - BEETLE (Zinn et al. 2002)
 - Geometry Explanation Tutor (Aleven et al. 2003)
 - **Why2/Atlas (VanLehn et al. 2002)**
 - students explain physical systems
 - **ITSpoke (Litman et al. 2006)**
 - builds upon Why2, spoken language based
 - SCOT (Pon-Barry et al. 2006)
 - ProPL (Lane and VanLehn 2005)
 - **AutoTutor (Graesser et al. 2003)**
 - students answer deep questions about computer technology
- a core set of foundational requirements for mixed-initiative natural language interaction in tutorial dialogue

17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 133/206

Corpus-Based Studies



- Speech acts in tutorial dialogue (Marineau et al. 2000)
- Dialogue acts' correlation with learning (Forbes-Riley et al. 2005, Core et al. 2003, Rosé et al. 2003, Katz et al. 2003)
- Student uncertainty in dialogue (Liscombe et al. 2005, Forbes-Riley and Litman 2005)
- Comparing text-based and spoken dialogue (Litman et al. 2006)

17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 134/206

Cognitive and Affective States in Learning



- ITS as platforms to investigate the impact of tutorial interactions on affective and motivational outcomes (e.g. self-efficacy) along with cognitive measures (i.e. learning gains)
- Goal: identifying tutorial strategies that balance the tradeoff between cognitive and affective learning outcomes
- Widespread methodology: investigate human-human tutorial dialogues (e.g. Boyer et al. 2008)

17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 135/206

ITS Interaction Style



- By dialogue initiative:
 - System initiative
 - Mixed-initiative
- By interaction modality:
 - Text-based
 - Speech-based

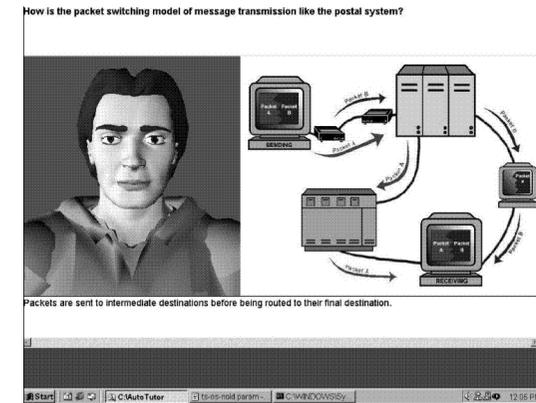
17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 136/206

AutoTutor



- Tutoring Research Group at the University of Memphis (e.g. Graesser et al., 1999)
- Intended for college students who take an introductory course in computer literacy
 - Fundamentals of computer hardware, operating system and the Internet
- Goals:
 - To comprehend student contributions
 - To simulate dialogue moves of normal (unskilled) or sophisticated tutors

Screenshot of AutoTutor (Graesser et al., 2001)



35

Interface Description



- Major problem is printed at the top of the screen
- Major questions are generated from a curriculum script:
 - Questions invite lengthy explanations and deep reasoning
 - *Why, how* and *what-if* questions
- Deep reasoning rather than short snippets of shallow knowledge
- 10 to 30 turns for a single question from a curriculum script
- Learner's contributions are typed in

Example Tutorial Dialogue (AutoTutor: Graesser et al., 2001)



TUTOR-1: Alright, let's go on. You want to take photos and send them over the Internet. Consider this problem: What hardware do you need to take photos and send them over the Internet?

STUDENT-1: First of all you need to have access to the Internet.

TUTOR-2: What else?

STUDENT-2: a camera

TUTOR-3: Yeah. For what type of connection do you need a network card?

STUDENT-3: network card

TUTOR-4: Okay. Now, if you have access to the Internet through a network card, then your connection is...

STUDENT-4: I don't know.

TUTOR-5: Is direct. Okay. A network card is needed if you have a direct connection to the Internet. Okay. How does the user get hooked up to the Internet?

STUDENT-5: by a plug.

TUTOR-6: Once again, a network card is needed if you have a direct connection to the Internet. Okay, what does a modem do?

STUDENT-6: connects network card and internet

TUTOR-7: Okay. Try this: A modem is required when there is...

STUDENT-7: a connection

TUTOR-8: a dial up connection.

...

TUTOR-30: Let's review. To send your photos on the Internet, you need either a digital camera or a regular camera to take the photos. If you use a regular camera, you need a scanner to scan them onto a computer disk. If you have a direct connection to the Internet, then you need a network card. A modem is needed if you have a dial up connection.

Information Delivery versus Knowledge Construction



- The answer is not graded (good / bad / score)
- Multi-turn conversation to extract more information from the student
- Students learn by constructing explanations and elaborations of the material (e.g. Chi et al., 1994)

System Architecture



1. Animated agent
 - Tree-dimensional
2. Curriculum script
 - Important concepts, questions, cases, and problems
3. Speech act classifier
 - Segmenting, parsing student's response, rule-based utterance classification
4. Latent semantic analysis (LSA)
 - Evaluating the quality of students' contributions
5. Dialogue move generator
 - Can include question answering, repeating the question, encouraging
6. Dialogue Advancer Network
 - Uses speech act and LSA to select next dialogue move and discourse marker

7. Question answering tool

How to Engage the Student in Conversation?



- Dialogue moves:
 - E.g. open-ended pumps, e.g. *What else?*
 - Tutors have a set of expectations about what to include into the answer
 - Expectation-1
 - Expectation-2
- AutoTutor decides what expectation to handle next and selects a dialogue move
 - Hints (indirect)
 - Prompts (in-between)
 - Assertions (direct)
- Exit the cycle when the student articulated the expected answer

How to Evaluate the Quality of the Answer?



- Match students utterances to expectations
- Statistical, corpus-based measure of representing knowledge
 - Latent Semantic Analysis (LSA)
- *max* function considering the current utterance and all combinations with previous learner's utterances
- An expectation is considered covered if it exceeds some threshold value

How to Select the Next Expectation to Cover?



- Use LSA in conjunction with various criteria
- Use next expectation with the highest score below threshold (zone of proximal development)
- Use next expectation with the highest LSA overlap with the previous covered expectation (coherence)
- Further constraints to advance the agenda in an optimal way

How to Give Feedback to a Student?



- Three channels of feedback:
 - **Backchannel** – acknowledge the learner's input, based on important nouns, e.g. *uh-huh*
 - **Pedagogical** feedback on the learner's previous turn, based on LSA scores
 - Negative, e.g. *not really*
 - Neutral negative, e.g. *okay*
 - Neutral positive, e.g. *okay*
 - Positive, e.g. *right*
 - **Corrective feedback** – repair bugs and misconceptions
 - Need to be explicitly anticipated

Dialogue Management



- Dialogue advancer network (DAN), mixed-initiative dialogue
- Formally an augmented state transition network
 - Selection of dialogue move on turn N+1 is sensitive a large set of parameters computed from dialogue history
- Student: What does X mean?
Tutor: answer by giving definition from a glossary
- Student: gives an assertion
Tutor: evaluate the quality and give short evaluative feedback

Types of Dialogue Moves



- Pump
- Hint
- Splice
- Prompt
- Prompt response
- Elaboration
- Summary
- Five forms of immediate short-feedback

Curriculum Script



- Organizes the content of topics covered in the dialogue
- Each topic is associated with:
 - A set of expectations
 - A set of hints and prompts for each expectation
 - A set of anticipated bugs/misconceptions and their corrections
 - (optionally) pictures or animations

Authoring Tools



- Create an LSA space
 - Identify a corpus of documents on the domain knowledge
- Lesson planner
 - Create a curriculum script with deep reasoning questions and problems
- Compute LSA vectors on the content of curriculum scripts
- Prepare glossary of important terms and their definitions

Domain Adaptation



Levels:

1. Glossary of terms and definitions (metacognition)
2. LSA space for conceptual physics (comprehension)
3. Curriculum script with deep reasoning questions and associated answers (production)
 - Most labour-intensive

Why2

(<http://www.pitt.edu/~vanlehn/why2000.html>)



- Chi et al. found that having students explain physical systems qualitatively positively correlated with learning outcomes
- Explanations can be done on formal and graphical languages, but also in natural languages
- Why2 targets to coach students explain physical systems in natural language
- Idea: ask the student to type in an explanation for a simple physical situation

Example dialogue

Question: A battery is connected to a bulb by two wires. The bulb lights. Why?



1. Student: There are two currents which come out, one current positive and one negative. They combine in the bulb to produce light and heat.
2. Tutor: I see. Which current is composed of moving electrons?
3. Student: I'm not sure. Both?
4. Tutor: Right, both are. An electrical current is just moving electrons. Which direction do the electrons move in your two currents?
5. Student: I think the electrons move out from the battery to the bulb. The two currents meet there and produce energy.
6. Tutor: Are the electrons actually destroyed in the bulb?
7. Student: I'm not sure.

8. Tutor: Actually, electrons are not destroyed or lost in an electrical current. Instead, they go out one side of the battery, go through the bulb and go back in the other side of the battery. Thus, your two currents are actually the same flow, only the electrons are leaving the battery for one and returning to the battery for the other. Given everything you know about batteries and bulbs, does this make sense?
9. Student: But I thought the battery dies when it runs out of electrons.
10. Tutor: Batteries die when they run out of the chemical energy that they use to pump electrons around the circuit. The battery is like a water pump that produces a powerful current, and the bulb is like a paddle wheel that is driven by the water current. When the water pump runs out of energy, the current slows to a stop. As this happens, does the number of electrons in the battery change?
11. Student: Maybe a little. But it stays about the same.
12. Tutor: Exactly right! I think you understand electrical current better now. See if you can write a better explanation.

Dialogue Management

- Student's utterance is analyzed to detect any misconceptions
- If a misconception is detected, a knowledge construction dialogue is initiated (KCD)
- Misconceptions are anticipated by collecting and analyzing a corpus of explanations from students

39

ITSpoke (Intelligent Tutoring SPOKEN dialogue system)

- A speech-enabled version of Why2-Atlas tutoring system
- Workflow:
 - The student's essay is parsed
 - A set of dialogue topics concerning misconceptions or incomplete explanations is extracted
 - ITSpoke then engages student in a dialogue that covers these topics
 - Therefore, the student revises the essay
 - End the tutoring problem
 - Cause another round of dialogue/essay revision

- Back-end is [Why2-Atlas](#) system (VanLehn et al. 2002)

- Back-end is [Why2-Atlas](#) system (VanLehn et al. 2002)

- Back-end is [Why2-Atlas](#) system (VanLehn et al. 2002)

System Architecture

- Sphinx speech recognizer (Huang et al., 1993)
 - Trained with example user utterances
 - Domain adaptation by human-computer typed corpus
 - Language model enhancement by human-human spoken language corpus
- Festival speech synthesizer (Black and Taylor, 1997)
- Sentence-level syntactic and semantic analysis modules (Rosé, 2000)
- Discourse and domain level processors (Makatchev et al., 2002)

ITSpoke Annotated Dialogue Excerpt

.....spoken dialogue excerpt at 6.1 minutes after start of problem...

ITSPOKE₄: Good. Does the packet have an acceleration? If yes please specify its direction. [ShortAnswerQ, Advance, Depth1, GoalRep0]

STUDENT₄: horizontally forward yes [Incorrect]

ITSPOKE₅: Well... To find the packet's acceleration we will use Newton's second law. What is the definition of this law? [LongAnswerQ, Push, Depth2, GoalRep0]

STUDENT₅: an object in motion tends to stay in motion until it's act by an outside force [Incorrect]

ITSPOKE₆: Well... Newton's second law says that the net force acting on an object equals the mass times the acceleration of the object. We can define Newton's second law as $F=m \cdot a$. Hence what can we determine about the acceleration of the packet given that we know the net force acting on it is not zero and points straight down? [DeepAnswerQ, Advance, Depth2, GoalRep0]

STUDENT₆: I don't know [Can't Answer]

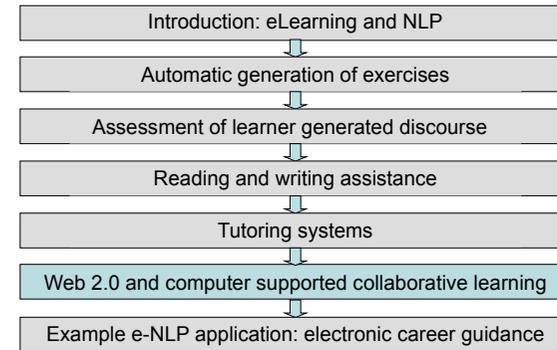
ITSPOKE₇: Well... Since the net force F and the acceleration a are vectors, for the equation $F=m \cdot a$ to hold they both need to have the same direction. <...more...> Does the packet have an acceleration? If yes please specify its direction. [ShortAnswerQ, PopUp, Depth1, GoalRep1]

Benefits of Spoken Interaction



- Benefits of human-human tutoring through spoken interaction (Lemke, 1990; Chi et al. 1994)
- Spontaneous self-explanation occurs more frequently in spoken tutoring (Hausmann and Chi, 2002)
- Speech contains prosodic and acoustic information to predict emotional states (Ang et al., 2002; Batliner et al., 2000)
 - Connection between learning and emotion (Coles, 1999)

Outline



41

Characteristics of Web 2.0

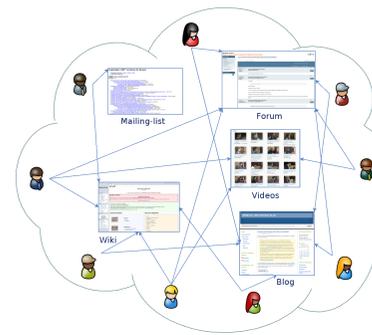


- Collective intelligence
- Huge amount of data
- Fast growing



- 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
 - ...

"CALL 2.0"



Select a language* Welcome guest Help Register Search Feedback Login

mocho BETA Home

Learn Sign up for a course. Start learning a new language. Email Password LOGIN

Practice Test your language skills. Break words or converse.

Share Tutor members learning your native language.

The social way to learn a language.

Community. Livemocha is the first-of-its-kind online language-learning community. Lessons. Fun and interactive lessons that move at the right pace for you. Motivation. Track your progress and reach your goals with Livemocha tools.

It's Free! Start now. It's Free! **Get Started** Or, learn more.

Discover Livemocha

Livemocha blends self-guided lessons, a vibrant community, and interactive tools to help you talk to the world.

Learn

- Enjoy fun self-study lessons.
- Build reading, listening, speaking, and writing skills.
- Build confidence to use what you learn.
- Learn three languages: English, French, and German, Spanish and Mandarin, Chinese.

Practice

- Practice conversing with text and audio tools.
- Receive one-on-one instruction from tutors and native speakers.
- Invite your friends to join the conversation and motivate each other.
- Connect anywhere and anytime.

Share

- Connect with a community of native speakers from around the world.
- Submit speaking and writing tips from native speakers.
- Return the favor! Help other language learners meet their goals.

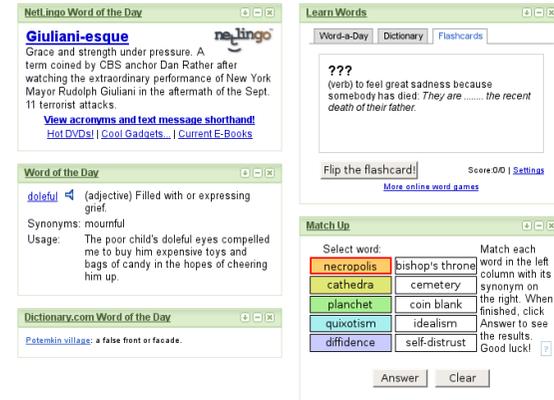
TAKE THE CHALLENGE

Meet our tutors

Access our team of awesome language tutors. They're available 9 days per week to chat one-on-one or to answer your questions.

Erik
Speaks: French, Italian, German, English

Widgets for CALL



NetLingo Word of the Day

Giuliani-esque

Grace and strength under pressure. A term coined by CBS anchor Dan Rather after watching the extraordinary performance of New York Mayor Rudolph Giuliani in the aftermath of the Sept. 11 terrorist attacks.

[View acronyms and text message shorthand!](#)
[Hot DVDs!](#) | [Cool Gadgets...](#) | [Current E-Books](#)

Word of the Day

doleful (adjective) Filled with or expressing grief.

Synonyms: mournful

Usage: The poor child's doleful eyes compelled me to buy him expensive toys and bags of candy in the hopes of cheering him up.

Dictionary.com Word of the Day

Patetina village: a false front or facade.

Learn Words

Word-a-Day Dictionary Flashcards

???
(verb) to feel great sadness because somebody has died. *They are the recent death of their father.*

Flip the flashcard! score: 00 | [Settings](#)
[More online word games](#)

Match Up

Select word: Match each word in the left column with its synonym on the right. When finished, click Answer to see the results. Good luck!

necropolis	bishop's throne
cathedra	cemetery
planchet	coin blank
quixotism	idealism
diffidence	self-distrust

Answer Clear

42

Use of Web 2.0 Resources



lingro

Translate Dictionary
Learn Web Videos
Collaborate Edit Videos
Lingo Jobs

Le Monde.fr

Rechercher

Le Monde

EN VENTE CHEZ VOTRE MARCHAND DE JOURNAU

Limagrain renonce à expérimenter ses maïs transgéniques en France

Alors que le feu vert des OGM revient devant l'Assemblée ce mardi, la grande entreprise fondée par des agriculteurs du Pays de Dôme renonce, cette année, à faire des essais de maïs transgénique en plein champ en France.

Valérie Péresse réforme le CNRS pour "décloisonner" la recherche

Le Centre national de la recherche scientifique va être transformé en institut national, dont sortira les sciences de la vie et les sciences de l'information.

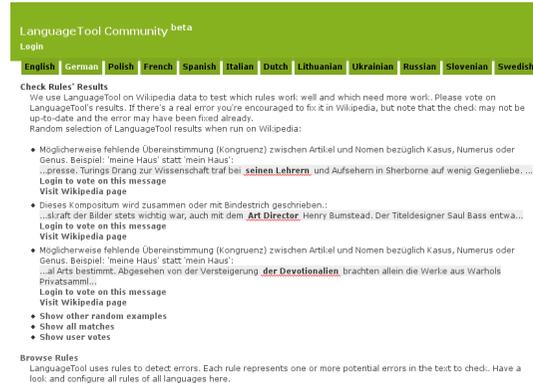
Wiktionary

[ˈwɪkʃənəri] n., a wiki-based Open Content dictionary

Wilen [ˈwɪlən kəʀnɪ]

Community-rule-based Grammar Checking

- A new paradigm? <http://community.languagetool.org>



LanguageTool Community beta

Login

English German Polish French Spanish Italian Dutch Lithuanian Ukrainian Russian Slovenian Swedish

Check Rules' Results

We use LanguageTool on Wikipedia data to test which rules work well and which need more work. Please vote on LanguageTool's results. If there's a real error you're encouraged to fix it in Wikipedia, but note that the check may not be up-to-date and the error may have been fixed already.

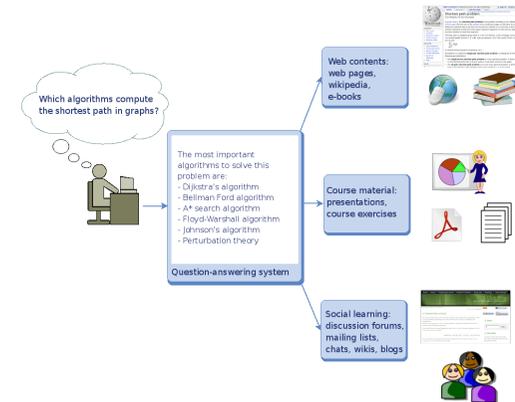
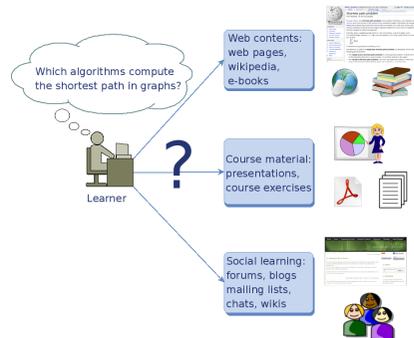
Random selection of LanguageTool results when run on Wikipedia:

- Möglicherweise fehlende Übereinstimmung (Kongruenz) zwischen Artikel und Nomen bezüglich Kasus, Numerus oder Genus. Beispiel: 'meine Haus' statt 'mein Haus' -> **grasak**. **Tunings** Orang zur Wissenschaft traf bei **seinen Lehrern** und Aufsehern in Sherborne auf wenig Gegenliebe...
Login to vote on this message
Visit Wikipedia page
- Dieses Kompositum wird zusammen oder mit Bindestrich geschrieben: ...schrift: der Bilder stets wichtig war, auch mit dem **Art Director**, Henry Bumstead. Der Titledesigner Saul Bass entwarf...
Login to vote on this message
Visit Wikipedia page
- Möglicherweise fehlende Übereinstimmung (Kongruenz) zwischen Artikel und Nomen bezüglich Kasus, Numerus oder Genus. Beispiel: 'meine Haus' statt 'mein Haus': ...Art bestimmt. Abgesehen von der Versteigerung **der Devotionalien**, brachten allein die Werke aus Warhols Privatammi...
Login to vote on this message
Visit Wikipedia page
- Show other random examples
- Show all matches
- Show user votes

Browse Rules

LanguageTool uses rules to detect errors. Each rule represents one or more potential errors in the text to check. Have a look and configure all rules of all languages here.

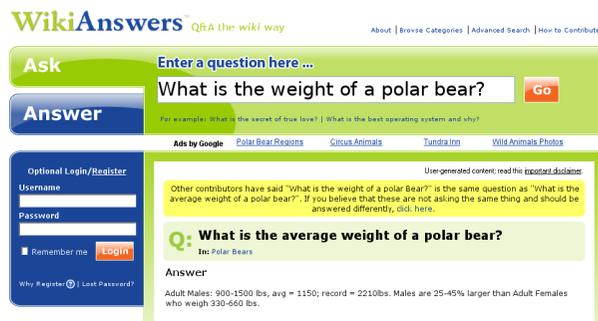
Motivation: Information overload in E-Learning



43

Social Q&A Sites

- Solution to the problem of automatically answering learners' questions: use repositories of already answered questions (Bernhard & Gurevych, 2008)



The screenshot shows the WikiAnswers website interface. At the top, it says "WikiAnswers™ Q&A the wiki way". Below that, there is an "Ask" section with a text input field containing "What is the weight of a polar bear?" and a "Go" button. Below the "Ask" section, there is an "Answer" section with a text input field containing "What is the average weight of a polar bear?". The answer section also includes a "Q:" field with the question "What is the average weight of a polar bear?" and an "Answer" field with the text "Adult Males: 900-1500 lbs, avg = 1150; record = 2210lbs. Males are 25-45% larger than Adult Females who weigh 320-660 lbs.".

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 resources:
 - 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 consent (Lampe and Resnick, 2004)



Case Study



Markus Weimer and Iryna Gurevych. 2007. **Predicting the Perceived Quality of Web Forum Posts**. RANLP, Borovetz, Bulgaria.

Goal: Develop a system to automatically assess the perceived quality of forum posts

44

Related Work on Quality Assessment



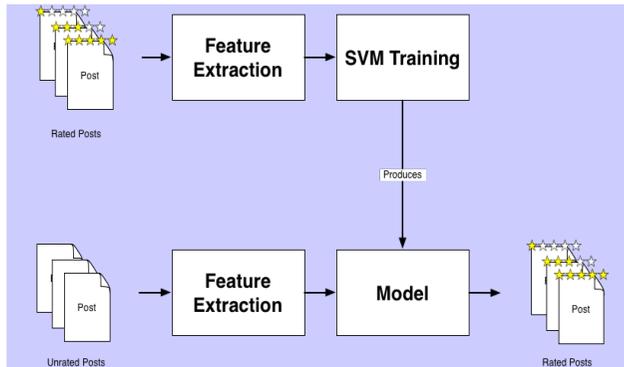
- **Essay scoring**
 - Established in systems like e-Rater (Attali and Burstein, 2006)
 - Very specialized approach: It is known what a “good” essay is
 - Input on which features to use
- **Automatically assessing review helpfulness** (Kim et al., 2006)
 - Goal: predict the helpfulness of product reviews on Amazon.com
 - Also very specialized:
 - The rating task is clearly defined: helpful / not helpful for buying decision
 - Dominant feature is metadata-dependent: star rating of the product

Requirements



- Adapt to the quality standards of a user community
- Be independent of metadata-based features
- Apply the system to forums from different domains

Approach in Weimer and Gurevych (2007)



Classification Features

- **Surface**
 - Length in tokens
 - Question Frequency
 - Exclamation Frequency
 - Capital WORD Frequency
- **Form Specific**
 - IsHTML
 - IsMail
 - Quote Fraction
 - URL Count
 - Path Count
- **Lexical**
 - Spelling Error Frequency
 - Swear Word Frequency
- **Syntactic**
 - Part of speech distribution
- **Similarity**
 - Cosine between the post unigram and the forum unigram

45

Data

- Provided by Nabble.com
- Preprocessing of the data:
 - Removal Non-English posts
 - Removal of posts with a rating of exactly 3 stars
 - Binarization of the data into good/bad posts
- Three data sets:
 - ALL: All the posts
 - SOFT: Posts from the software category at Nabble.com
 - MISC: Posts from the other categories
- Data available upon request

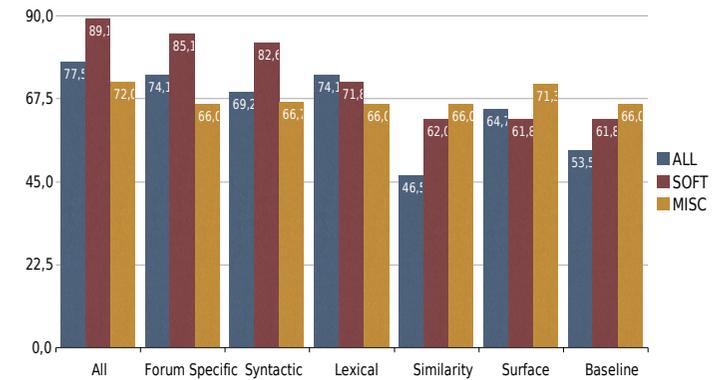
Descriptive Statistics

	ALL	SOFT	MISC
Unfiltered Posts	4291	1968	2323
All ratings three stars	135 3%	61 3%	74 3%
Contradictory ratings	70 2%	14 1%	56 2%
No text	56 1%	30 2%	26 1%
Non-English	668 15%	361 18%	307 13%
Remaining	3418 80%	1532 78%	1886 81%
Good Posts	1829 54%	947 62%	1244 66%
Bad Posts	1589 46%	585 38%	642 34%

Experiments: Setup

- Stratified tenfold cross validation with different feature sets
- Evaluation measure: mean average precision
- Features were extracted using Apache UIMA
- Classifier:
 - LibSVM
 - Gauss Kernel
 - Parameters $C = 10$, $\gamma = 0.1$
 - No model selection was performed
- Baseline: Majority class classifier

Results



46

Error Analysis: Confusion Matrix

	true good	true bad	sum	
pred. good	1517	456	1973	ALL
pred. bad	312	1133	1445	
sum	1829	1589	3418	

	true good	true bad	sum	
pred. good	490	72	562	SOFT
pred. bad	95	875	970	
sum	585	947	1532	

	true good	true bad	sum	
pred. good	1231	516	1747	MISC
pred. bad	13	126	139	
sum	1244	642	1886	

Error Analysis: Typical Errors

- **Automatically generated mails**
 - Can be filtered out in preprocessing
- **Non-textual content**
 - May be used as a feature, e.g. code examples in a software developer's forum
- **Very short posts**
 - Might be improved through metadata about the user or thread information
- **Opinion based ratings**
- **Ratings based on domain knowledge**
 - Probably form the upper bound for our approach

Ratings Based on Domain Knowledge



> Thank You for the fast response, but I'm not sure if I understand you right. INTERRUPTS can be interrupted (by other interrupts or signals) and SIGNALS not.

Yup. And I responded faster than my brain could shift gears and got my INTERRUPT and SIGNAL crossed.

> All my questions still remain!

Believe J"org addressed everything in full. That the compiler simply can't know that other routines have left zero reg alone and the compiler expects to find zero there.

As for SREG, no telling what another routine was doing with the status bits so it too has to be saved and restored before any of its contents possibly get modified. CISC CPUs do this for you when stacking the IRQ, and on RTI.

Human rating: -
System rating: +

Opinion Based Ratings



> But you would impose US law even in a country where smoking weed is legal

Given that most of our users and most significant press coverage is American, yes. That is why I drew the line there.

Yes, I know it isn't perfect. But it's better than anything else I've seen.

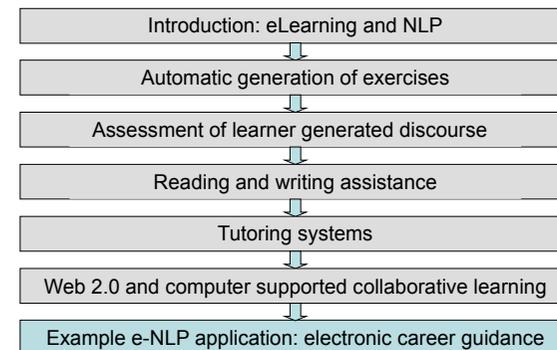
Human rating: -
System rating: +

Conclusions



- Quality assessment is machine learnable
- The system performs best with forum specific features (~90%)
- Even without forum specific features, the system gives satisfactory result (~82%)
- Further experiments needed on:
 - different data sets
 - types of user-generated discourse
- New classification features:
 - structure of the forum
 - lexical semantic features

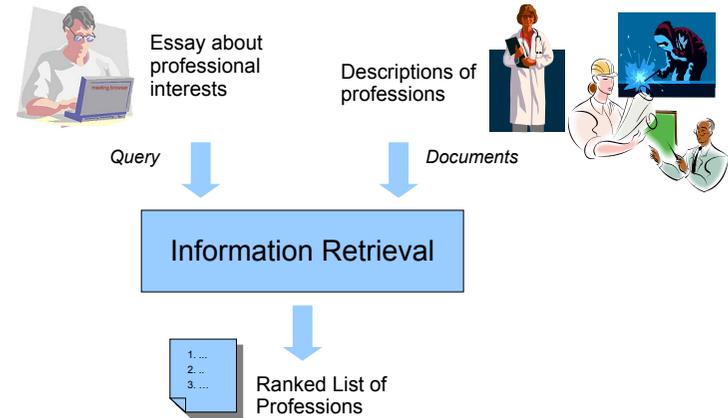
Outline



The SIR project: Semantic Information Retrieval for Electronic Career Guidance

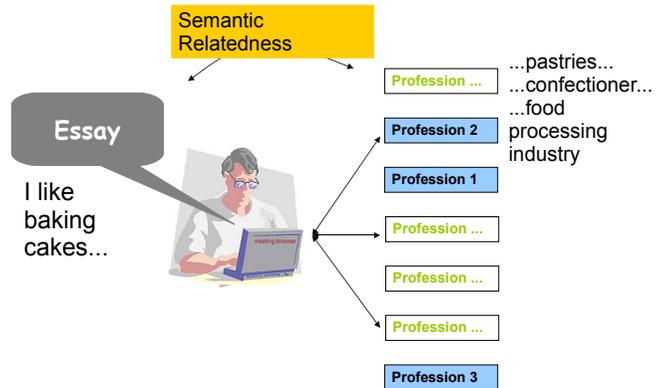
Deutsche Forschungsgemeinschaft funded by the German Research Foundation
DFG

Electronic Career Guidance



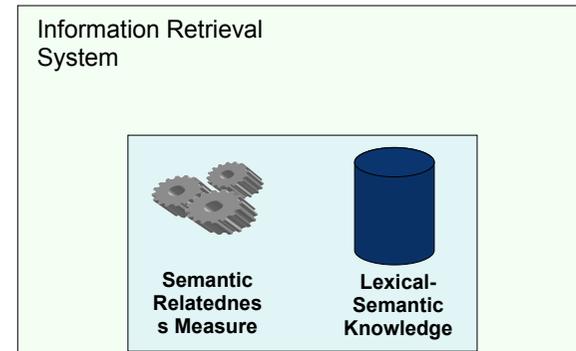
48

Vocabulary Mismatch Problem



Semantic IR Models

- Semantic relatedness (SR) as measure for document relevance



Lexical Semantic Knowledge



- 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>

17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 193/206

Semantic Relatedness Measures



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

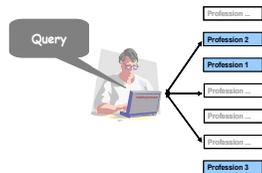
17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 194/206

49

Experiments in Information Retrieval



“Andererseits arbeite ich besonders gerne am Computer, kann programmieren in C, Python und VB und könnte mir deshalb auch vorstellen in der Software-Entwicklung zu arbeiten.”



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

17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 195/206

Document Collection



- 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

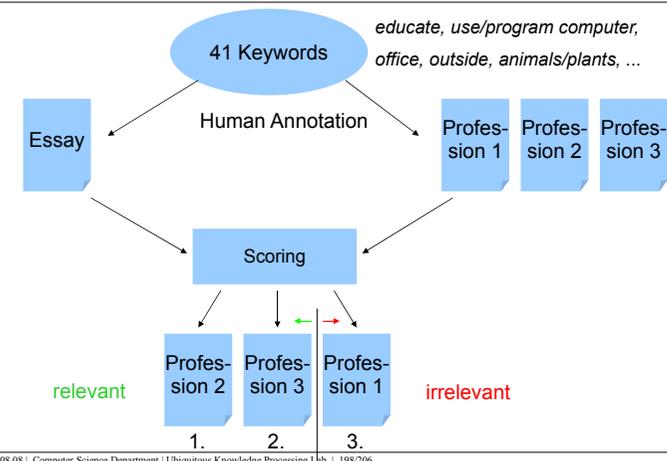
17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 196/206

"Gold Standard"



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

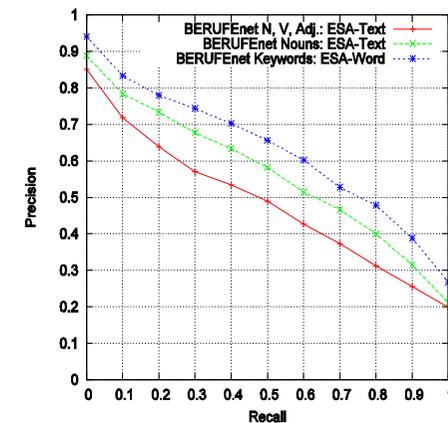
Relevance Judgments



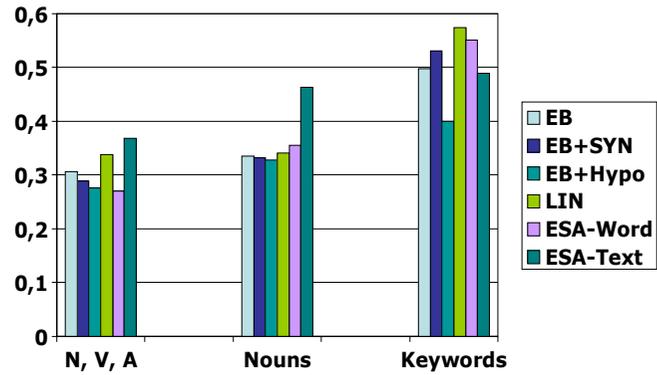
Evaluation

- Standard IR measures using relevance judgements
 - Precision – recall diagrams
 - Mean average precision
- Rank correlation with knowledge-based ranked list
 - Spearman's Rank Correlation Coefficient
- Parameters:
 - Pre-processing configurations
 - Semantic relatedness measures
 - Lexical-semantic knowledge sources

Pre-processing Configurations & Measures, Precision-Recall

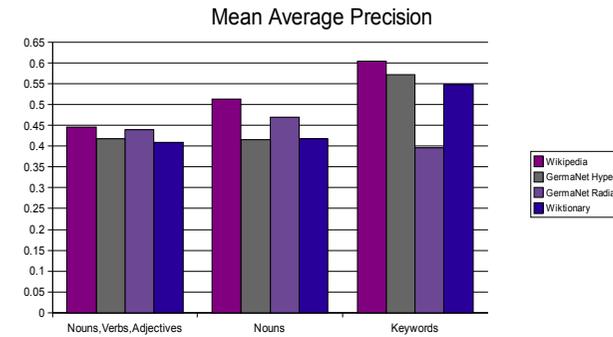


Pre-processing Configurations & Measures Spearman's Rank Correlation



17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 201/206

ESA-Text tf.idf with Different Lexical-Semantic Resources



17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 202/206

Some Thoughts on eLearning 2.0...



- Opportunity for NLP and e-NLP?
 - Remove knowledge acquisition bottleneck
 - New forms of eLearning
- Excellent playground for NLP?
 - eLearning 2.0 discourse types almost not studied
- Can we actually learn from BioNLP?

17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 203/206

How to Promote e-NLP?



- Establish an international community
- ACL-associated meeting series (e.g. ACL-BEA Workshop 2008)
- Related Tutorials
- Resources:
 - Bibliography
 - Research groups
 - Projects
 - Annotated corpora
 - Tools

17.08.08 | Computer Science Department | Ubiquitous Knowledge Processing Lab | 204/206

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
 - Analyzing collaborative learning processes automatically
 - Learner's corpora and resources
 - eLearning standards, e.g. SCORM

Thank you!
<http://www.ukp.tu-darmstadt.de/>

