Natural Language Processing
SoSe 2014

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Introduction to Language Technology

Dr. Mariana Neves

April 16th, 2014

Outline

- NLP course
- Introduction to NLP
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- Course materials

Outline

- NLP course
- Introduction to NLP
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- Course materials

NLP Course

- Home page:
  - https://epic.hpi.uni-potsdam.de/Home/NaturalLanguageProcessingSS2014
- Lecture
  - Wednesday 13:30-15:00
  - HS 3
  - 3 credit points
- Assessment
  - Attendance to 80% of the sessions (8/11 sessions)
  - Deliver of the exercises
  - Final exam
  - 40% (exercises), 60% (exam)
- Contact
  - Mariana.Neves@hpi.uni-potsdam.de
  - Room 1.02 (Villa), Tuesday 13:00-14:00 or under request

NLP Course

- Acknowledgments:
  - Dr. Saeehed Momtazi (slides)

- Remarks:
  - Many examples on the biomedical domain, but no previous knowledge needed

Outline

- NLP course
- Introduction to NLP
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- Course materials

Types of Language

- Natural languages
  - English
  - German
  - Japanese
- Programming languages
  - Java
  - Python

Natural Language

- A vocabulary consists of a set of words (w_i)
- A text is composed of a sequence of words from a vocabulary
- A language is constructed of a set of all possible texts

Examples of vocabulary

- English
  - the
  - eat
  - you
  - book

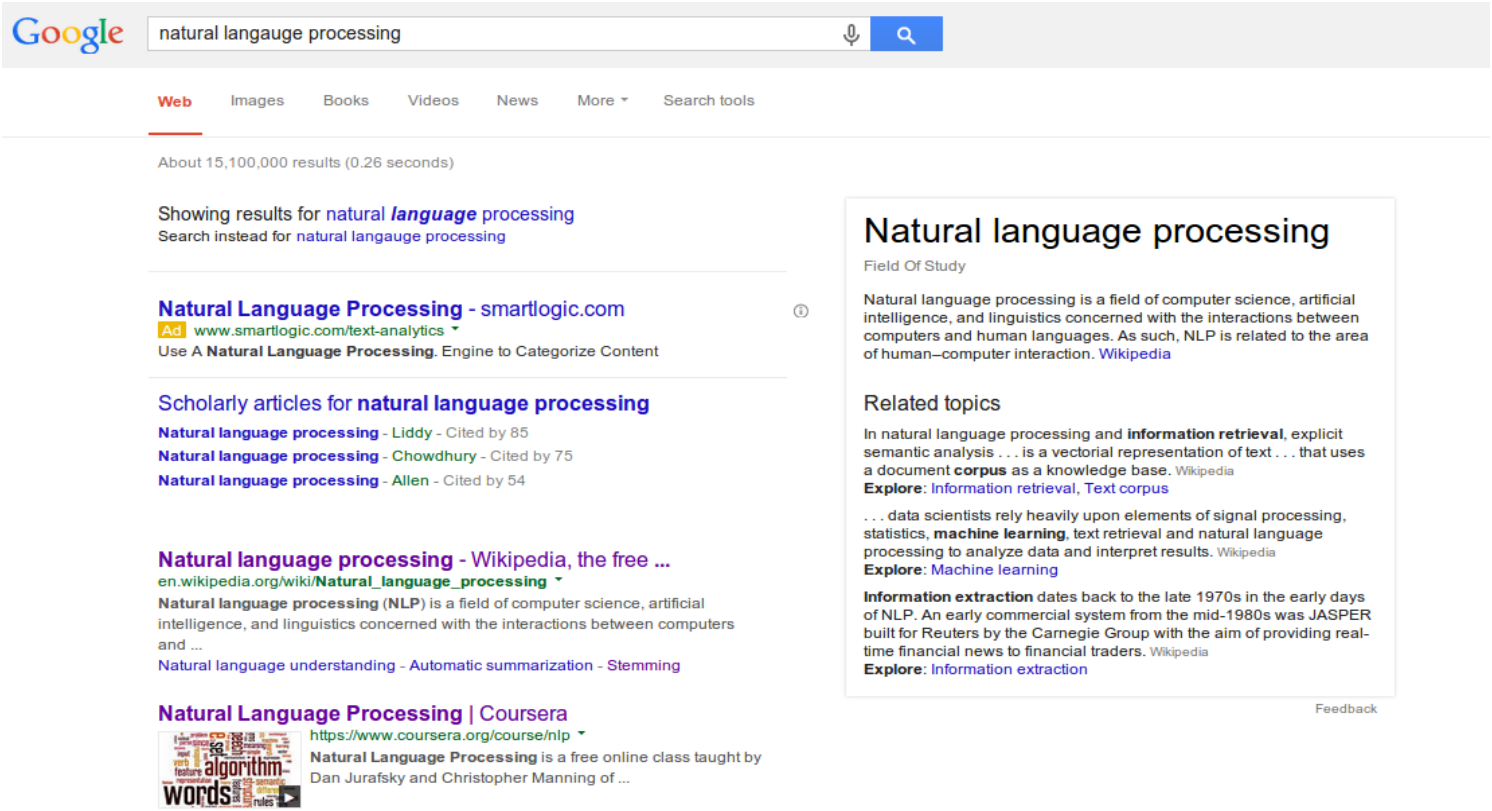
- German
  - das
  - essen
  - du
  - Buch

Outline

- NLP course
- Introduction to NLP
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- Course materials

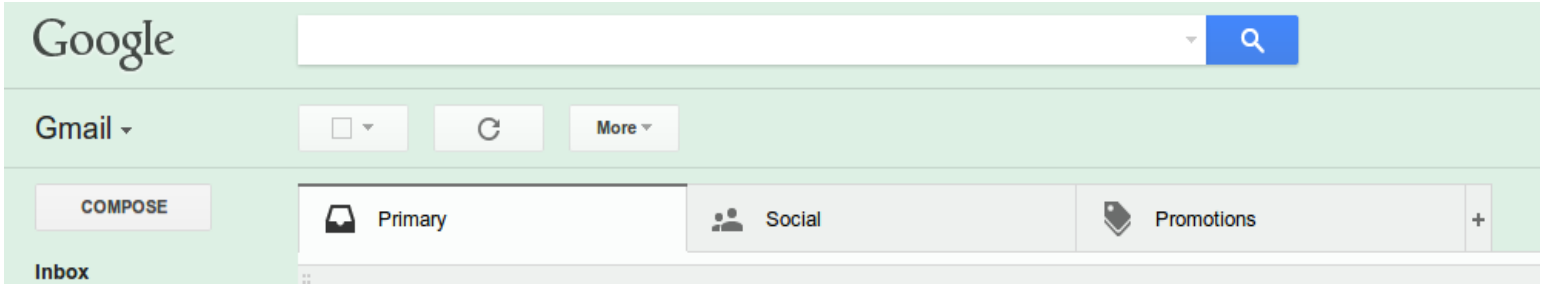
Spell and Grammar Checking

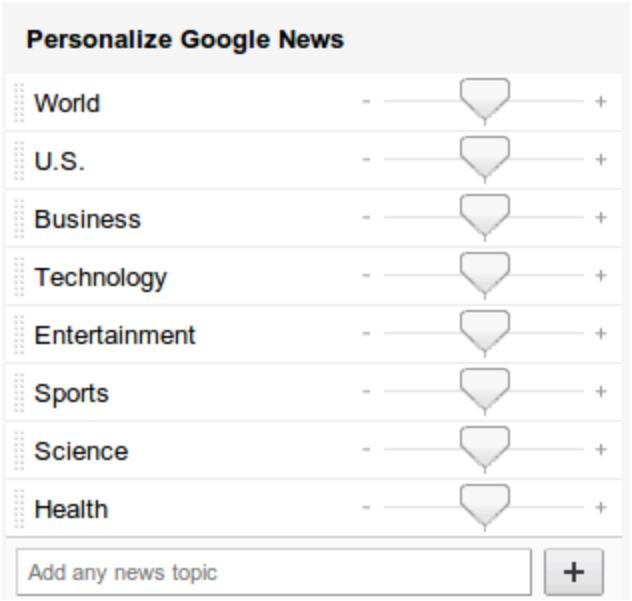
- Checking spelling and grammar of a text, and suggesting alternatives for the errors



Text Categorization

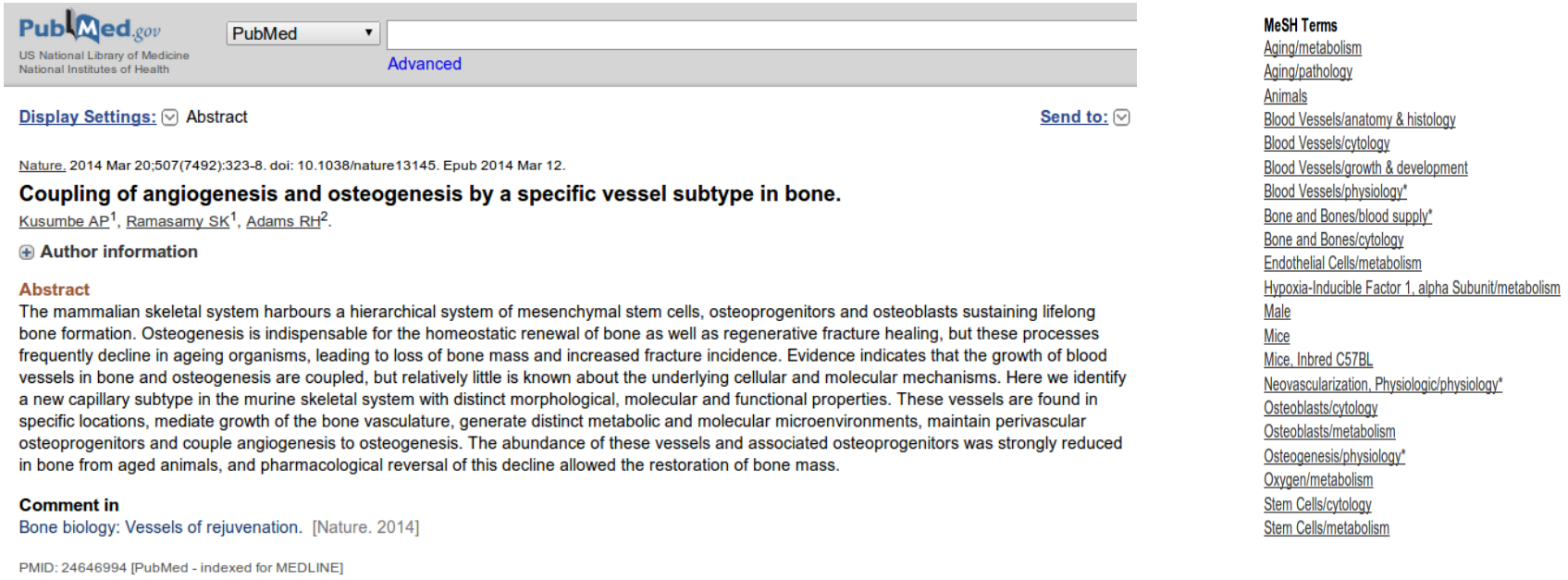
- Assigning a pre-defined category to a text





Text Categorization

- Assigning a pre-defined category to a text



PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed ▾ Advanced

Display Settings: Abstract Send to:

Nature. 2014 Mar 20;507(7492):323-8. doi: 10.1038/nature13145. Epub 2014 Mar 12.

Coupling of angiogenesis and osteogenesis by a specific vessel subtype in bone.

Kusumbe AP[1], Ramasamy SK[1], Adams RH[2].

Author information

Abstract

The mammalian skeletal system harbours a hierarchical system of mesenchymal stem cells, osteoprogenitors and osteoblasts sustaining lifelong bone formation. Osteogenesis is indispensable for the homeostatic renewal of bone as well as regenerative fracture healing, but these processes frequently decline in ageing organisms, leading to loss of bone mass and increased fracture incidence. Evidence indicates that the growth of blood vessels in bone and osteogenesis are coupled, but relatively little is known about the underlying cellular and molecular mechanisms. Here we identify a new capillary subtype in the murine skeletal system with distinct morphological, molecular and functional properties. These vessels are found in specific locations, mediate growth of the bone vasculature, generate distinct metabolic and molecular microenvironments, maintain perivascular osteoprogenitors and couple angiogenesis to osteogenesis. The abundance of these vessels and associated osteoprogenitors was strongly reduced in bone from aged animals, and pharmacological reversal of this decline allowed the restoration of bone mass.

Comment in

Bone biology: Vessels of rejuvenation. [Nature. 2014]

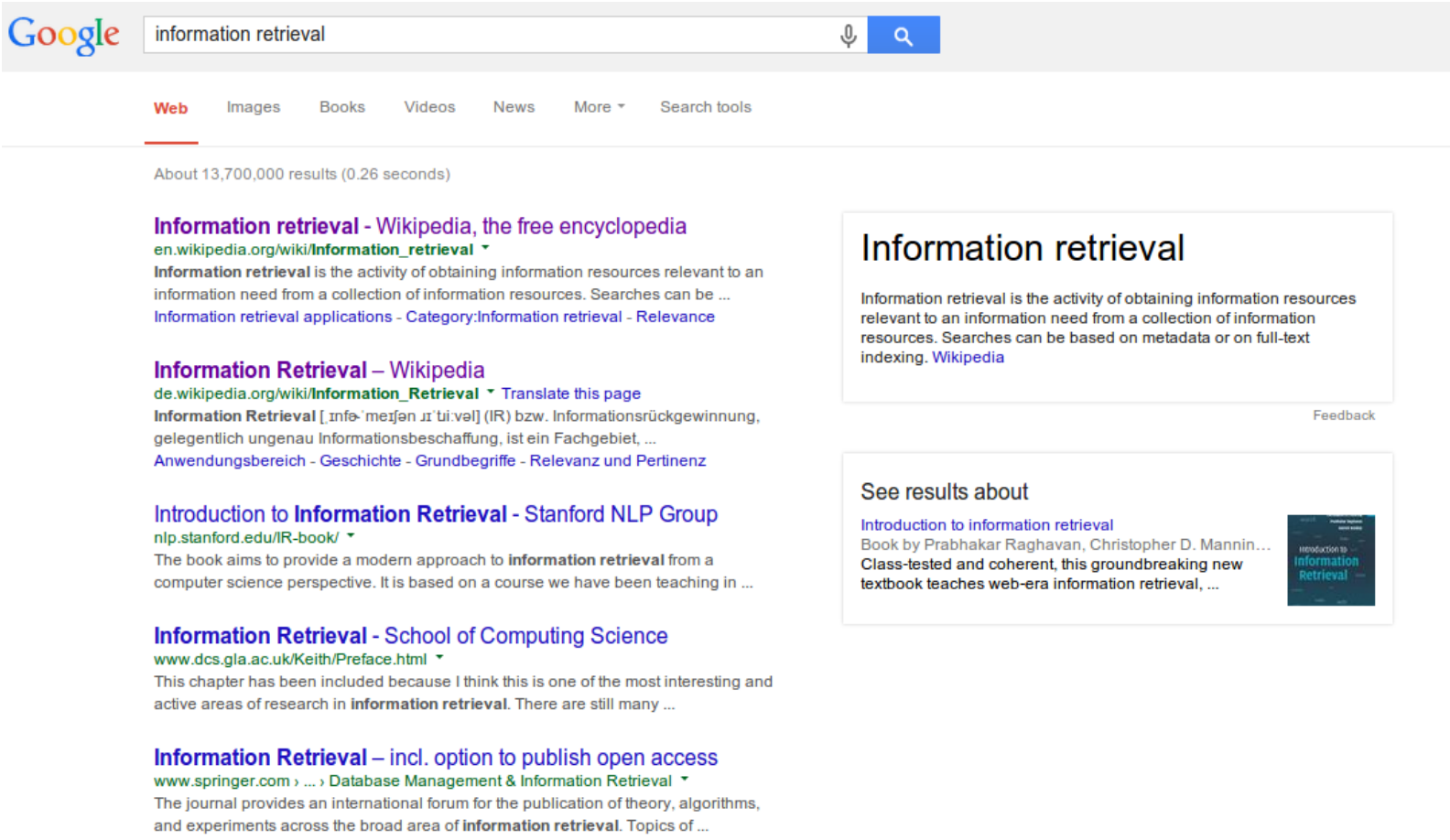
PMID: 24646994 [PubMed - indexed for MEDLINE]

MeSH Terms

Aging/metabolism
Aging/pathology
Animals
Blood Vessels/anatomy & histology
Blood Vessels/cytology
Blood Vessels/growth & development
Blood Vessels/physiology*
Bone and Bones/blood supply*
Bone and Bones/cytology
Endothelial Cells/metabolism
Hypoxia-Inducible Factor 1, alpha Subunit/metabolism
Male
Mice
Mice, Inbred C57BL
Neovascularization, Physiologic/physiology*
Osteoblasts/cytology
Osteoblasts/metabolism
Osteogenesis/physiology*
Oxygen/metabolism
Stem Cells/cytology
Stem Cells/metabolism

Information Retrieval

- Finding relevant information to the user's query



Summarization

- Generating a short summary from one or more documents, sometimes based on a given query

This is the companion website for the following book.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.

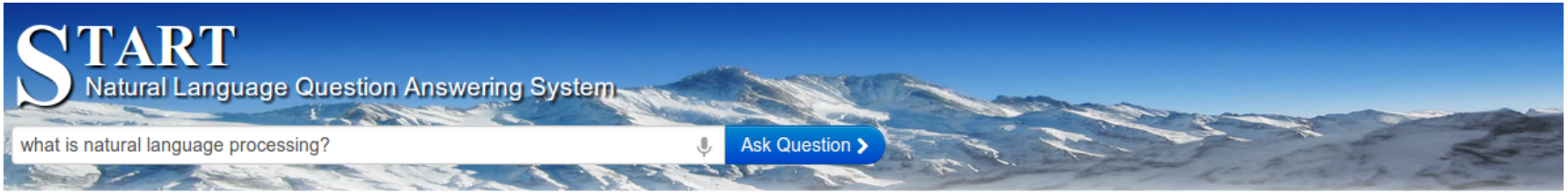
You can order this book at CUP, at your local bookstore or on the internet. The best search term to use is the ISBN: 0521865719.

The book aims to provide a modern approach to information retrieval from a computer science perspective. It is based on a course we have been teaching in various forms at Stanford University, the University of Stuttgart and the University of Munich.

We'd be pleased to get feedback about how this book works out as a textbook, what is missing, or covered in too much detail, or what is simply wrong. Please send any feedback or comments to: informationretrieval (at) yahoogroups (dot) com

Summarization

- Generating a short summary from one or more documents, sometimes based on a given query



===> what is natural language processing?

Natural language processing

Natural language processing (**NLP**) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation.

Source: Wikipedia

Summarization

- Generating a short summary from one or more documents, sometimes based on a given query

General annotation (Comments)

Function	Acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type. Involved in cell cycle regulation as a trans-activator that acts to negatively regulate cell division by controlling a set of genes required for this process. One of the activated genes is an inhibitor of cyclin-dependent kinases. Apoptosis induction seems to be mediated either by stimulation of BAX and FAS antigen expression, or by repression of Bcl-2 expression. In cooperation with mitochondrial PPIF is involved in activating oxidative stress-induced necrosis; the function is largely independent of transcription. Induces the transcription of long intergenic non-coding RNA p21 (lincRNA-p21) and lincRNA-Mkln1. LincRNA-p21 participates in TP53-dependent transcriptional repression leading to apoptosis and seem to have to effect on cell-cycle regulation. Implicated in Notch signaling cross-over. Prevents CDK7 kinase activity when associated to CAK complex in response to DNA damage, thus stopping cell cycle progression. Isoform 2 enhances the transactivation activity of isoform 1 from some but not all TP53-inducible promoters. Isoform 4 suppresses transactivation activity and impairs growth suppression mediated by isoform 1. Isoform 7 inhibits isoform 1-mediatedapoptosis. Ref.34 Ref.42 Ref.61 Ref.66 Ref.70 Ref.93 Ref.95 Ref.107 Ref.110 Ref.122 Ref.125
Cofactor	Binds 1 zinc ion per subunit.
Subunit structure	Interacts with AXIN1. Probably part of a complex consisting of TP53, HIPK2 and AXIN1 By similarity. Binds DNA as a homotetramer. Interacts with histone acetyltransferases EP300 and methyltransferases HRMT1L2 and CARM1, and recruits them to promoters. In vitro, the interaction of TP53 with cancer-associated/HPV (E6) viral proteins leads to ubiquitination and degradation of TP53 giving a possible model for cell growth regulation. This complex formation requires an additional factor, E6-AP, which stably associates with TP53 in the presence of E6. Interacts (via C-terminus) with TAF1; when TAF1 is part of the TFIID complex. Interacts with ING4; this interaction may be indirect. Found in a complex with CABLES1 and TP73. Interacts with HIPK1, HIPK2, and TP53INP1. Interacts with WWOX. May interact with HCV core protein. Interacts with USP7 and SYVN1. Interacts with HSP90AB1. Interacts with CHD8; leading to recruit histone H1 and prevent transactivation activity By similarity. Interacts with ARMC10, BANP, CDKN2AIP, NUAK1, STK11/LKB1, UHRF2 and E4F1. Interacts with YWHAZ; the interaction enhances TP53 transcriptional activity. Phosphorylation of YWHAZ on 'Ser-58' inhibits this interaction. Interacts (via DNA-binding domain) with MAML1 (via N-terminus). Interacts with MKRN1. Interacts with PML (via C-terminus). Interacts with MDM2; leading to ubiquitination and proteasomal degradation of TP53. Directly interacts with FBXO42; leading to ubiquitination and degradation of TP53. Interacts (phosphorylated at Ser-15 by ATM) with the phosphatase PP2A-PPP2R5C holoenzyme; regulates stress-induced TP53-dependent inhibition of cell proliferation. Interacts with PPP2R2A. Interacts with AURKA, DAXX, BRD7 and TRIM24. Interacts (when monomethylated at Lys-382) with L3MBTL1. Isoform 1 interacts with isoform 2 and with isoform 4. Interacts with GRK5. Binds to the CAK complex (CDK7, cyclin H and MAT1) in response to DNA damage. Interacts with CDK5 in neurons. Interacts with AURKB, SETD2, UHRF2 and NOC2L. Interacts (via N-terminus) with PTK2/FAK1; this promotes ubiquitination by MDM2. Interacts with PTK2B/PYK2; this promotes ubiquitination by MDM2. Interacts with PRKCG. Interacts with PPIF; the association implicates preferentially tetrameric TP53, is induced by oxidative stress and is impaired by cyclosporin A (CsA). Interacts with human cytomegalovirus/HHV-5 protein UL123. Interacts with SNAI1; the interaction induces SNAI1 degradation via MDM2-mediated ubiquitination and inhibits SNAI1-induced cell invasion. Interacts with KAT6A. Interacts with UBC9. Interacts with ZNF385B; the interaction is direct. Interacts (via DNA-binding domain) with ZNF385A; the interaction is direct and enhances p53/TP53 transactivation functions on cell-cycle arrest target genes, resulting in growth arrest. Interacts with ANKRD2. Interacts with RFFL (via RING-type zinc finger); involved in p53/TP53 ubiquitination. Ref.8 Ref.34 Ref.38 Ref.42 Ref.43 Ref.54 Ref.55 Ref.56 Ref.57 Ref.58 Ref.59 Ref.61 Ref.62 Ref.64 Ref.65 Ref.66 Ref.67 Ref.68 Ref.72 Ref.73 Ref.74 Ref.75 Ref.76 Ref.78 Ref.80 Ref.81 Ref.83 Ref.86 Ref.87 Ref.88 Ref.89 Ref.92 Ref.93 Ref.94 Ref.99 Ref.101 Ref.103 Ref.105 Ref.106 Ref.107 Ref.112 Ref.113 Ref.116 Ref.117 Ref.119 Ref.121 Ref.122 Ref.124 Ref.125 Ref.126 Ref.127 Ref.129 Ref.137 Ref.138 Ref.139 Ref.140 Ref.141 Ref.151

Information Extraction

- Extracting the important items of a text and assigning them a slot in a certain structure

Hasso Plattner Institute

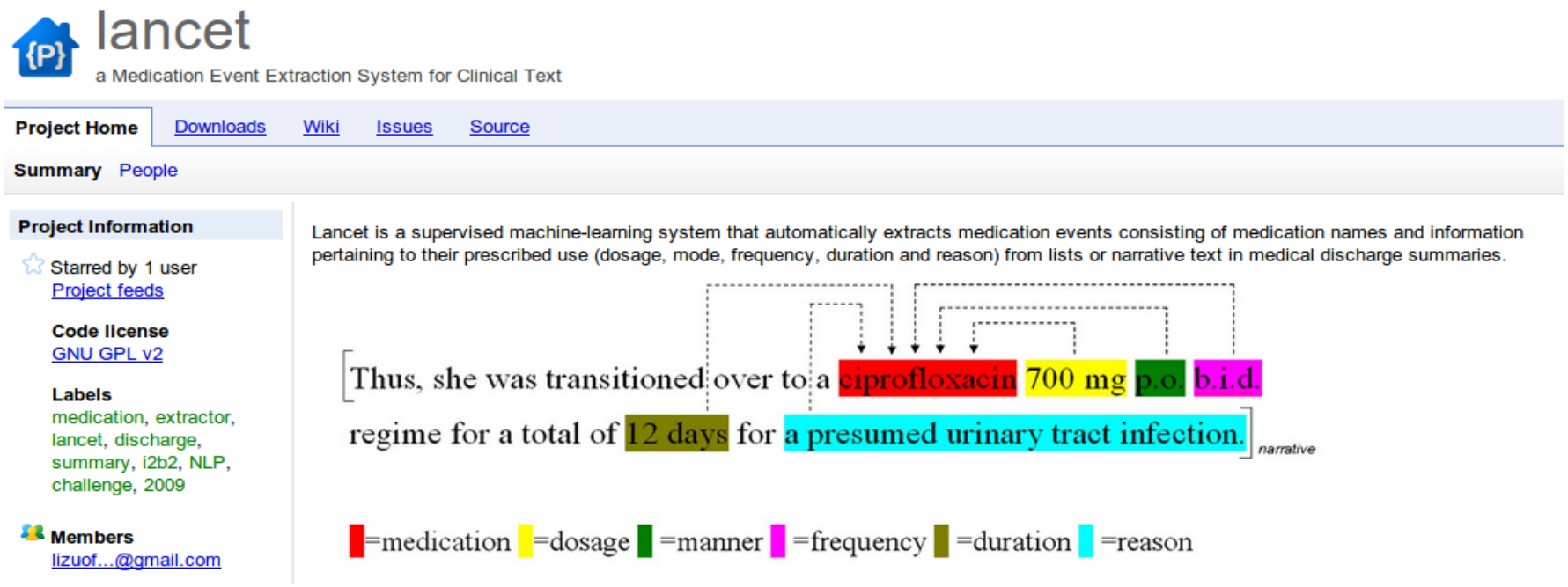
Hasso-Plattner-Institut für Softwaresystemtechnik GmbH

IT Systems Engineering | Universität Potsdam

Motto	Design IT. Create Knowledge.
Established	1998[1]
Type	Private university institute
Director	Prof. Dr. Christoph Meinel
Admin. staff	60[2]
Students	about 480[2]
Location	Potsdam, Germany
Campus	Griebnitzsee
Colors	Orange Vivid orange Dark pink
Affiliations	University of Potsdam
Website	www.hpi.uni-potsdam.de

Information Extraction

- Extracting the important items of a text and assigning them a slot in a certain structure



Question answering

- Answering questions asked by the user with a short answer

===> What cities are within 250 miles of the capital of Italy?

I know that the capital of Italy is Rome, Italy (source: START KB).

Using this information, I determined what cities are within 250 miles of the Rome, Italy:

In Italy, the following cities are within 250 miles of Rome:

Naples, Italy is 118 miles (189.90298 kilometers) from Rome.

Florence, Italy is 143 miles (230.13666 kilometers) from Rome.

Pisa, Italy is 148 miles (238.18338 kilometers) from Rome.

Bologna, Italy is 178 miles (286.4638 kilometers) from Rome.

Venice, Italy is 237 miles (381.4153 kilometers) from Rome.

Trieste, Italy is 241 miles (387.85266 kilometers) from Rome.

Verona, Italy is 244 miles (392.68073 kilometers) from Rome.

Genoa, Italy is 249 miles (400.72745 kilometers) from Rome.

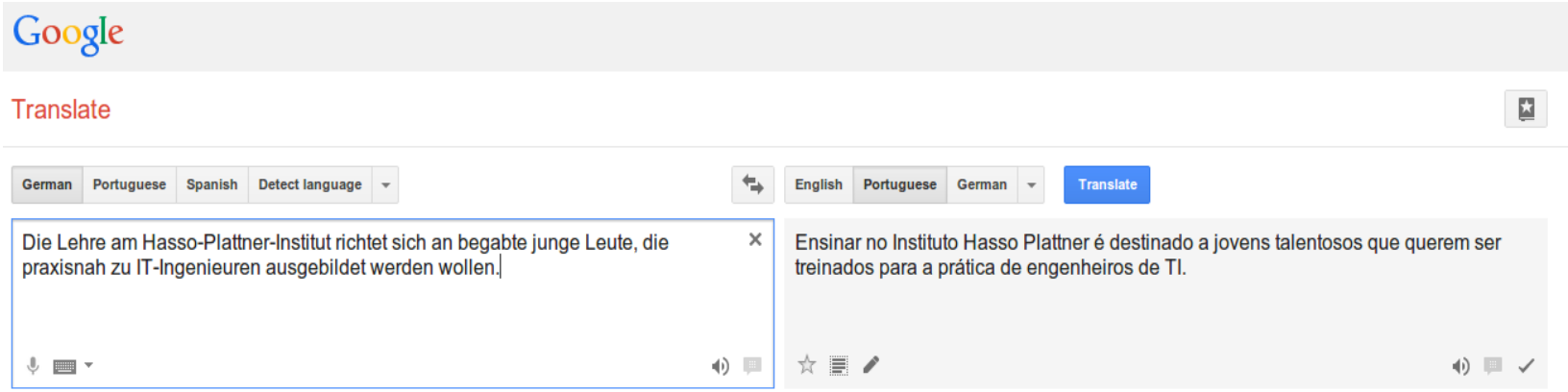
Source: START KB

Question answering

- Answering questions asked by the user with a short answer

Machine Translation

- Translating a text from one language to another



FOLHA DE S.PAULO

★ ★ ★ UM JORNAL A SERVIÇO DO BRASIL

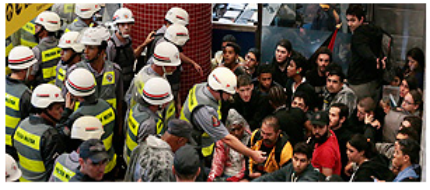
WEDNESDAY, APRIL 16, 2014 05:01

Dilma advertising spending totaled R $ 2.3 billion in 2013 and hit record

Plateau attributes the increase to campaigns such as anti-crack and the More Doctors

Group suggests that the anti-crisis Sabesp still use less water Cantareira

Board also wants discussion on 'constraint uses of water resources'



ATO AGAINST THE GLASS

After protest, four people remain detained in Sao Paulo

According to police, two teenagers and two young men were caught breaking banks

HESITATION

Andrew Vargas says now is 'reestudando' resignation to the mandate

Reuters/Yonhap

Ao menos uma pessoa morreu em naufrágio de embarcação com mais de 470 pessoas na Coreia do Sul; resgate prossegue na região

reader panel

COMMENT

TELEVISION

Reader laments end of the display TV program on Leaf Culture

Contact Sheet

EMPREENDEDOR SOCIAL 2014

columnists

FULL LIST

Delfim Net

'Noise' around the search Datafolha is exaggerated

Penny

It's hard to be a great coach, everything is uncertain

Marcelo Coelho

Cup songs are not always the national tone

Shuttle

Brazil raises export of wine and sparkling

Sentiment Analysis

- Identifying sentiments and opinions stated in a text

Customer Reviews

Speech and Language Processing, 2nd Edition

15 Reviews

5 star: (8)
4 star: (3)
3 star: (3)
2 star: (0)
1 star: (1)

Average Customer Review
(15 customer reviews)

Share your thoughts with other customers

Create your own review

The most helpful favorable review

4 of 4 people found the following review helpful

Great introductions and reference book
I read the first edition of that book and it is terrific. The second edition is much more adapted to current research. Statistical methods in NLP are more detailed and some syntax-based approaches are presented. My specific interest is in machine translation and dialogue systems. Both chapters are extensively rewritten and much more elaborated. I believe this book is...
Read the full review ›
Published on August 9, 2008 by carheg

› See more **5 star**, **4 star** reviews

Vs.

The most helpful critical review

37 of 37 people found the following review helpful

Good description of the problems in the field, but look elsewhere for practical solutions
The authors have the challenge of covering a vast area, and they do a good job of highlighting the hard problems within individual sub-fields, such as machine translation. The availability of an accompanying Web site is a strong plus, as is the extensive bibliography, which also includes links to freely available software and resources.

Now for the...
Read the full review ›
Published on April 2, 2009 by P. Nadkarni

› See more **3 star**, 2 star, **1 star** reviews

Sentiment Analysis

- Identifying sentiments and opinions stated in a text

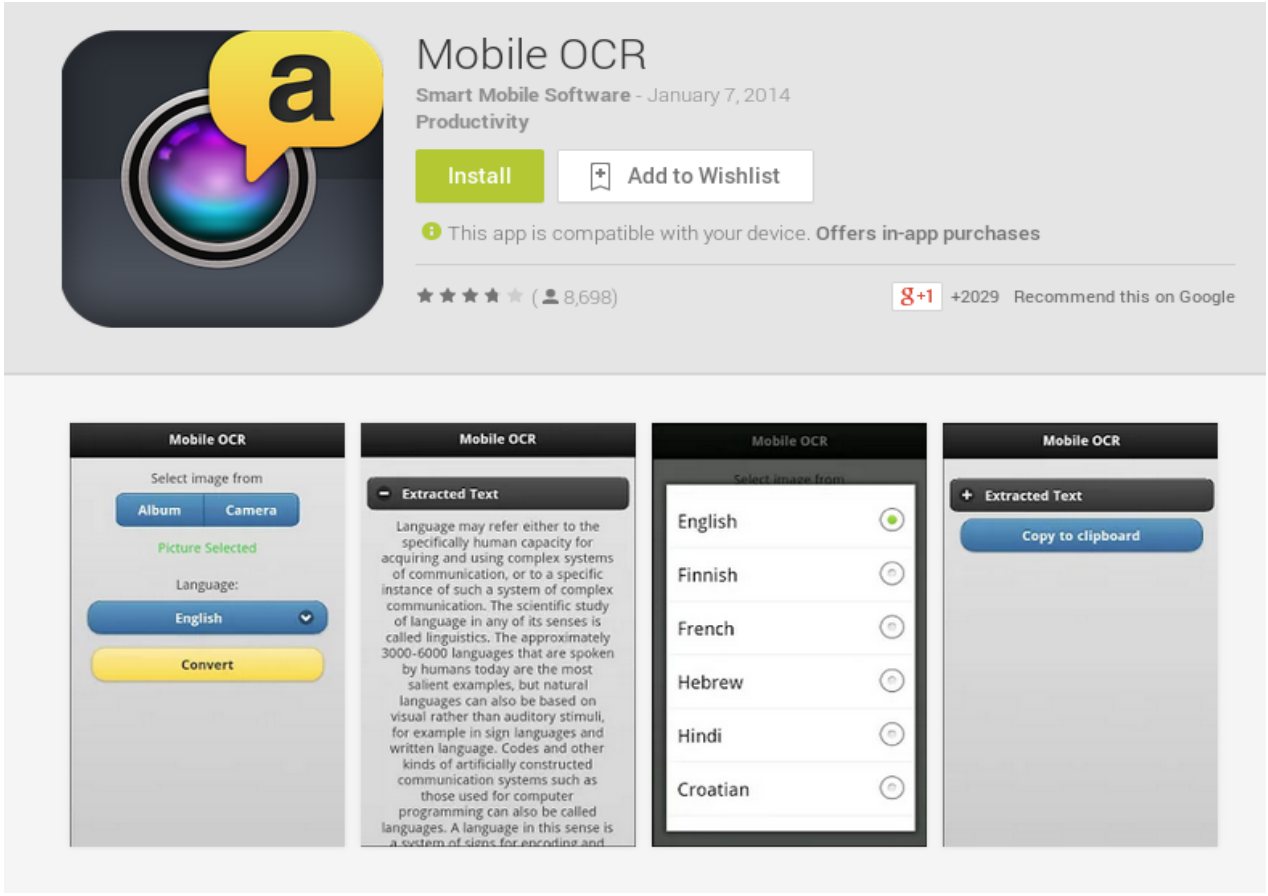
SemEval-2014 Task 9

Task Description: Sentiment Analysis in Twitter

Authorities are *only too aware* that Kashgar is 4,000 kilometres (2,500 miles) from Beijing but *only* a tenth of the distance from the Pakistani border, and are *desperate* to *ensure instability or militancy* does not leak over the frontiers.
Taiwan-made products *stood a good chance* of becoming *even more competitive thanks to* wider access to overseas markets and lower costs for material imports, he said.
"March *appears* to be a *more reasonable* estimate while earlier admission *cannot be entirely ruled out*," according to Chen, also Taiwan's chief WTO negotiator.
friday evening plans were great, but saturday's plans *didnt go as expected* – i went dancing & it was an *ok* club, but *terribly crowded :-(*
WHY THE *HELL* DO YOU GUYS ALL HAVE MRS. KENNEDY! SHES A FUCKING DOUCHE
AT&T was *okay* but whenever they do something *nice* in the name of customer service it seems like a favor, while T-Mobile makes that a *normal everyday thin*
obama should be *impeached* on *TREASON* charges. Our Nuclear arsenal was TOP Secret. Till HE told our enemies what we had. *#Coward #Traitor*
My graduation speech: "I'd like to *thanks* Google, Wikipedia and my computer! *:D* #iThingteens

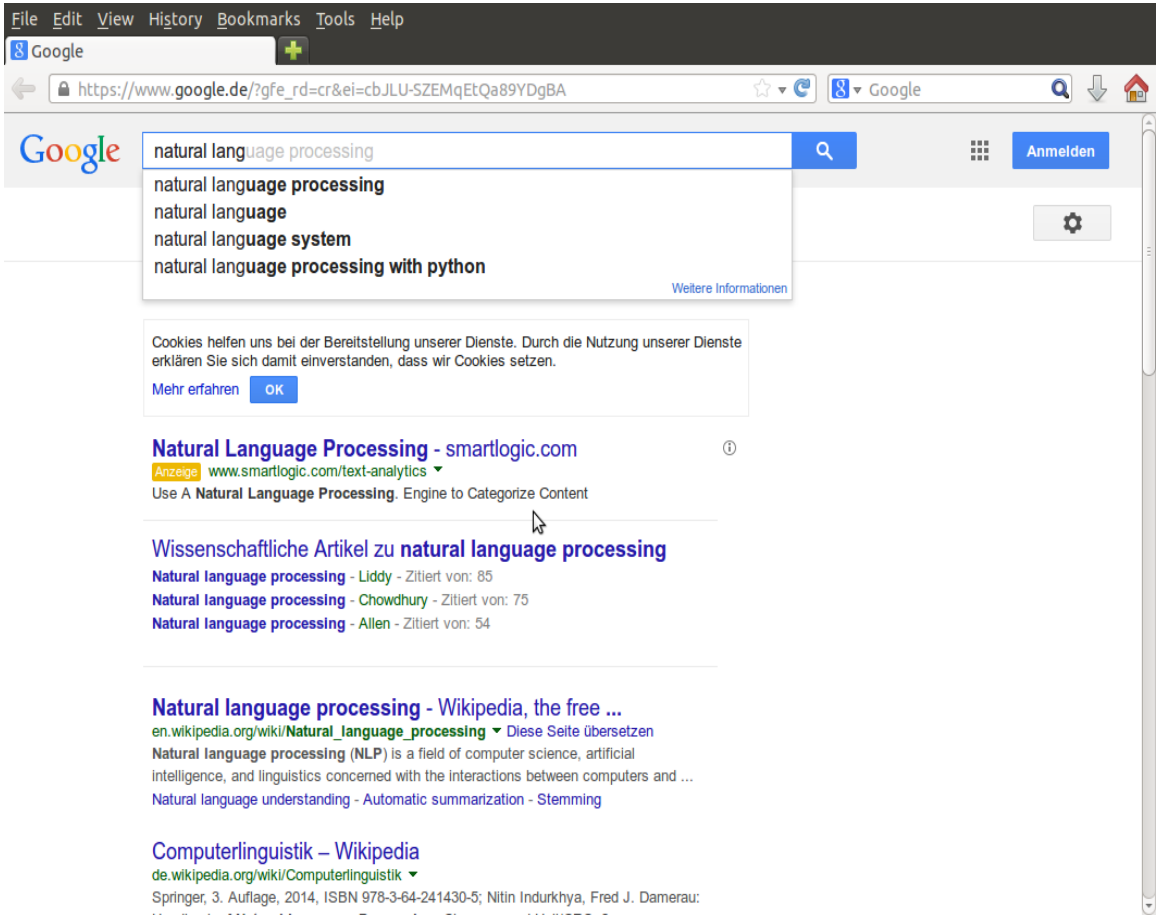
Optical Character Recognition

- Recognizing printed or handwritten texts and converting them to computer-readable texts



Word Prediction

- Predicting the next word that is highly probable to be typed by the user



Speech recognition

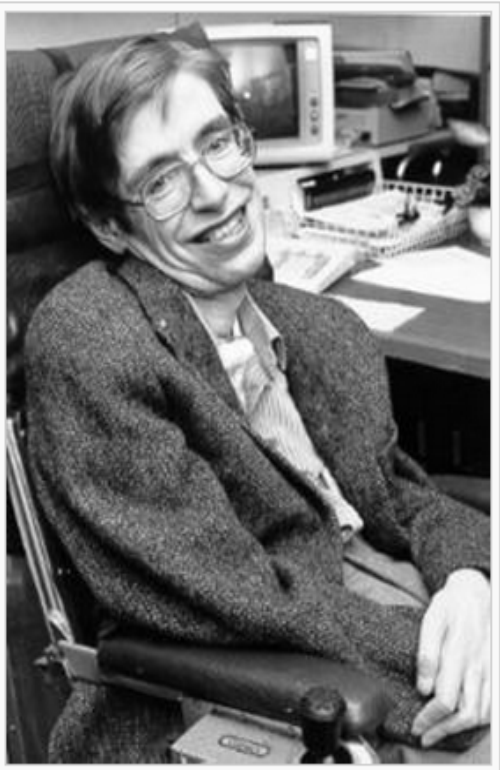
- Recognizing a spoken language and transforming it into a text

Siri.
Your wish is
its command.

Siri lets you use your voice to send messages, schedule meetings, place phone calls, and more. Ask Siri to do things just by talking the way you talk. Siri understands what you say, knows what you mean, and even talks back. Siri is so easy to use and does so much, you'll keep finding more and more ways to use it.

Speech synthesis

- Producing a spoken language from a text



Stephen Hawking is one of the most famous people using speech synthesis to communicate

Spoken dialog systems

- Running a dialog between the user and the system

Siri.
Your wish is
its command.

Siri lets you use your voice to send messages, schedule meetings, place phone calls, and more. Ask Siri to do things just by talking the way you talk. Siri understands what you say, knows what you mean, and even talks back. Siri is so easy to use and does so much, you'll keep finding more and more ways to use it.

Level of difficulties

- Easy (mostly solved)
  - Spell and grammar checking
  - Some text categorization tasks
  - Some named-entity recognition tasks

- Intermediate (good progress)
  - Information retrieval
  - Sentiment analysis
  - Machine translation
  - Information extraction

Level of difficulties

- Difficult (still hard)
  - Question answering
  - Summarization
  - Dialog systems

Outline

- NLP course
- Introduction to NLP
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- Course materials

Section splitting

- Splitting a text into sections

Eur Radiol
DOI 10.1007/s00330-014-3135-8

BREAST

Correlation between three-dimensional ultrasound features and pathological prognostic factors in breast cancer

Jun Jiang · Ya-qing Chen · Yi-zhuan Xu · Ming-li Chen · Yun-kai Zhu · Wen-bin Guan · Xiao-jin Wang

Received: 13 November 2013 / Revised: 30 January 2014 / Accepted: 17 February 2014

Abstract

Objectives To investigate the correlation of three-dimensional (3D) ultrasound features with prognostic factors in invasive ductal carcinoma.

Methods Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included. Morphology features and vascularization perfusion on 3D ultrasound were evaluated. Pathologic prognostic factors, including tumour size, histological grade, lymph node status, oestrogen and progesterone receptor status (ER, PR), c-erbB-2 and p53 expression, and microvessel density (MVD) were determined. Correlations of 3D ultrasound features and prognostic factors were analysed.

Results The retraction pattern in the coronal plane had a significant value as an independent predictor of a small tumour size (P=0.014), a lower histological grade (P=0.009) and positive ER or PR expression status (P=0.001, 0.044). The retraction pattern with a hyperechoic ring only existed in low-grade and ER-positive tumours. The presence of the hyperechoic ring strengthened the ability of the retraction pattern to predict a good prognosis of breast cancer. The increased intra-tumour vascularization index (VI, the mean tumour vascularity) reflected a higher histological grade (P=0.025) and had a positive correlation with MVD (r=0.530, P=0,001).

Conclusions The retraction pattern and histogram indices of VI provided by 3D ultrasound may be useful in predicting prognostic information about breast cancer.

Key Points

- *Three-dimensional ultrasound can potentially provide prognostic evaluation of breast cancer.*
- *The retraction pattern and hyperechoic ring in the coronal plane suggest good prognosis.*
- *The increased intra-tumour vascularization index reflects a higher histological grade.*
- *The intra-tumour vascularization index is positively correlated with microvessel density.*

Keywords Breast · Neoplasms · Ultrasound · Three-dimensional · Prognostic factors

Introduction

The three strongest prognostic factors in invasive breast cancer are widely accepted to be the size of tumour, histological grade and lymph node stage. The larger tumour size (>2 cm), high nuclear grade, and lymph node-positive status usually predict the aggressive biological behaviour with a high recurrence rate and a low survival rate. In addition, the tumour size and lymph node status greatly influence the choice of operative procedure and the decision to administer neoadjuvant chemotherapy [1, 2].

Biological markers such as oestrogen receptors (ER), progesterone receptors (PR), human epidermal growth factor receptor 2 (c-erbB-2) and the p53 index can also be used for prediction of medical treatment response and patient prognosis. The presence of ER and PR in breast cancer always determines the application of antihormonal therapy and usually indicates a good prognosis. Expression of c-erbB-2 or the p53 index is a powerful and independent prognostic factor for lymph node metastasis and tumour infiltration [1, 3]. Microvessel density (MVD) is the current reference standard in the characterization of tumour angiogenesis and has been shown to be associated with tumour growth, invasion, metastasis and disease-specific survival [4].

Three-dimensional (3D) ultrasound can afford additional information such as morphology features on the coronal plane and a global appearance of the mass vascularity, which cannot be achieved with conventional ultrasound. Therefore, it has been increasingly considered as an important imaging modality for evaluating primary breast cancer. However, so far, 3D ultrasound has been used mainly to differentiate benign and malignant lesions; no reports address correlations between the 3D ultrasound features and prognostic factors [5–7]. We therefore investigated possible correlation between the 3D ultrasound characteristics of invasive ductal carcinoma with pathologic prognostic factors to determine whether 3D ultrasound could be useful in the non-invasive prognostic evaluation of breast cancer.

Materials and methods

Patients

This retrospective study was approved by the ethical standards of the institutional ethics committee, and informed consent was obtained from all patients.

From September 2011 to May 2013, 85 patients with 85 lesions, pathologically proven to be invasive ductal carcinoma, were included in this study. The exclusion criteria were pregnancy or lactation, administration of preoperative chemotherapies or adjuvant chemotherapies. Patients with a breast mass larger than 3.0 cm were also excluded because more than one 3D volume acquisition was necessary to include the whole lesion plus 3 mm surrounding the breast lesion. All patients were female and aged 26 to 90 years (mean age, 56.3 years).

Ultrasound examination

All ultrasound images were obtained with one type of system (GE Voluson E8 Expert, Zipf, Austria) by two radiologists with 7–12 years of experience in breast ultrasound. An 11 L-D linear transducer with a frequency of 5–12 MHz was used for 2D ultrasound, and an RSP6-16-D dedicated volume transducer with a frequency of 6–12 MHz was used for 3D ultrasound.

Ultrasound examination was performed with patients in the supine position with elevated arms. Once the breast lesion was detected and the region of interest had been identified, the volume box was superimposed and set to include the entire display screen so as to cover the lesion and maximum amount of normal surrounding tissue. The sweep angle was adjusted to 15–29° according to the size of the breast lesion. Then the ultrasound probe was held still with enough jelly to contact the skin gently. The volume mode was switched on and the 3D ultrasound volume was generated by the automatic rotation of the mechanical transducer. When the first ultrasound examination was finished, the power Doppler mode was added for the second examination and the fixed preinstalled power Doppler settings used were 0.3 kHz pulse repetition frequency, "low 1" wall motion filter, −2.0 gain and high frequency. The first examination for 3D greyscale imaging took 10–20 s and the second, for 3D power Doppler imaging, took 25–45 s, depending on the size of the tumour. Then the total acquisition time for 3D ultrasound was about 1–2 min. The entire examination was saved in DICOM format and stored on the hard disk for further analysis.

Image analysis

The 3D ultrasound images were reviewed for this analysis by another two radiologists with 8–10 years of experience in breast ultrasound and characterized by consensus. In addition, the radiologists had not performed the data acquisition and were blinded to the patients' clinical and mammographic findings.

The ultrasound image was opened by using the 4D View software. Firstly, the tomographic ultrasound imaging (TUI) was used for a slice by slice documentation in the coronal plane. Then, the volume contrast imaging (VCI) and the surface render mode were added for better observation of the lesion and the surrounding tissue. All the slices were carefully observed to identify the presence of the retraction pattern in the surrounding tissue and the margin of the lesion. The retraction pattern was defined as the hyperechoic straight lines that radiated perpendicularly from the surface of the solid nodule, producing a stellar pattern [8, 9] (Fig. 1). The presence of the retraction pattern was further divided into with or without a hyperechoic ring, which was displayed as an echogenic halo ring between the mass and the surrounding tissue in the coronal plane (Fig. 2a).

The 3D power Doppler imaging analyses were performed using a virtual organ computer-aided analysis (VOCAL)-imaging program (GE, Zipf, Austria), which could automatically calculate the histogram indices of vascularization index (VI), flow index (FI) and vascularization flow index (VFI). VI represents the vessels in the defined volume by measuring the number of colour voxels in the region of interest, i.e. the mean tumour vascularity; FI represents the average intensity of flow by measuring the mean colour value in the colour voxels, i.e. the mean blood flow volume; VFI represents both regression modelling techniques to identify the most significant and independent 3D image findings. A P value less than 0.05 was considered statistically significant.

Results

Prognostic factors

In the current study group, the surgical specimens revealed 75 lesions with pure invasive ductal carcinoma and the remaining 10 lesions with invasive ductal carcinoma with DCIS components. The mean percentage of the DCIS components in the lesion was 8.10±4.93 % (range, 2–20 %).

The size of 85 lesions ranged from 5 to 30 mm, and the mean size was 19.92 mm (SD = 7.56 mm). Of the 85 tumours, 47 (55.3 %) were equal to or smaller than 2 cm and 38 (44.7 %) were larger than 2 cm. According to the Elston–Ellis grading system, there were 58 (68.2 %) grade II tumours and 27 (31.8 %) grade III. Lymph node metastasis was present in 30 (35.3 %) patients. There were 58 (68.2 %) ER-positive, 54 (63.5 %) PR-positive, 70 (82.4 %) c-erbB-2-positive and 42 (49.4 %) p53-positive tumours.

Table 1 Association between MVD and prognostic factors

Prognostic factor	*N*	Mean	SD	*P* value
Tumour size (cm)				
≤2	47	19.30	5.25	
>2	38	25.60	7.60	0.007
Tumour grade				
I/II	58	19.83	5.55	
III	27	25.83	8.02	0.023
Lymph node				
Negative	55	21.31	6.70	
Positive	30	22.08	7.34	0.946
ER				
Negative	27	23.27	8.36	
Positive	58	20.93	5.14	0.931
PR				
Negative	31	25.00	8.59	
Positive	54	19.82	5.09	0.092
c-erbB-2				
Negative	15	21.50	9.57	
Positive	70	21.55	6.65	0.788
p53				
Negative	43	23.13	7.04	
Positive	42	19.63	6.20	0.083

Correlation between MVD and prognostic factors

Significantly higher MVD was observed in the larger size group (P<0.01) and higher grade group (P<0.05). There were no significant associations between MVD and other pathological factors (P>0.05) (Table 1).

Correlation between morphological features and prognostic factors

Of the 85 breast lesions, 57 (67.1 %) showed the retraction pattern in the coronal plane of 3D ultrasound. Of these 57 lesions, 17 (29.8 %) showed the retraction pattern with a hyperechoic ring and 40 (70.2 %) were without the hyperechoic ring.

The tumour size, histological grade, ER and PR status all showed significant associations with the presence of the retraction pattern (P<0.01) (Table 2). Tumours with the retraction pattern were significantly more likely to be small in size, low grade, ER-positive and PR-positive (Fig. 3). Moreover, the retraction pattern with a hyperechoic ring, which presented as intricately mixed fibrous tissues and infiltrating carcinoma cells on pathological specimens, only existed in low-grade and ER-positive tumours (Fig. 2). The odds ratios of tumour size, tumour grade, and ER and PR status for patients with the retraction pattern and a hyperechoic ring versus no retraction pattern were all higher than those with the retraction pattern without a hyperechoic ring versus no retraction pattern (Table 3). The presence of the hyperechoic ring strengthened the ability of the retraction pattern to predict these good prognoses. However, the lymph node status and the expression of c-erbB-2 and p53 showed no statistically significant correlation with the retraction pattern (P>0.05).

As for MVD, however, no significant correlation was found between MVD and the presence of the retraction pattern on 3D ultrasound (P>0.05).

Correlation between vascularization perfusion and prognostic factors

For intra-tumoral regions, the mean VI, FI and VFI of 85 lesions were 6.84 (range, 0.02–21.61), 37.72 (range, 21.81–53.32) and 2.64 (range, 0.04–9.11), respectively. For shells with a thickness of 3 mm surrounding the breast lesion, the VI, FI and VFI were 7.31 (range, 0.14–25.13), 38.72 (range, 23.27–56.90) and 2.88 (range, 0.04–11.08), respectively.

Compared with the small tumours, the tumour foci with a diameter greater than 2 cm were more likely to show a higher inVI, inFI, inVFI, out3mmVI and out3mmVFI. The tumours with a high grade or lymph node metastasis had a higher inVI, inVFI, out3mmVI and out3mmVFI than the tumours with low grade or lymph node-negative status. ER-negative tumours had a higher inFI than ER-positive tumours and the tumours with negative expression of PR had a higher inVI, inVFI and out3mmVFI than PR-positive tumours (Table 4).

J. Jiang · Y.-q. Chen (✉) · Y.-z. Xu · M.-l. Chen · Y.-k. Zhu
Department of Ultrasound, Xinhua Hospital Affiliated to Shanghai Jiaotong University School of Medicine, 1665 Kongjiang Road, Shanghai 200092, China
e-mail: joychen_1266@163.com

W.-b. Guan
Department of Pathology, Xinhua Hospital Affiliated to Shanghai Jiaotong University School of Medicine, 1665 Kongjiang Road, Shanghai 200092, China

X.-j. Wang
Teaching and Research Section of Statistics, Shanghai Jiaotong University School of Medicine, 227 Chongqing South Road, Shanghai 200025, China

Published online: 12 April 2014

Sentence splitting

- Splitting a text into sentences

11 Sentences (= "T-" or "Terminable" units *only* if independent clauses are puctuated as separate sentences, e.g. "I came and he went"-->"I came. And he went.")
Average 23.55 words (SD=12.10)

OBJECTIVES: To investigate the correlation of three-dimensional (3D) ultrasound features with prognostic factors in invasive ductal carcinoma.

METHODS: Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included.

Morphology features and vascularization perfusion on 3D ultrasound were evaluated.

Pathologic prognostic factors, including tumour size, histological grade, lymph node status, oestrogen and progesterone receptor status (ER, PR), c erbB-2 and p53 expression, and microvessel density (MVD) were determined.

Correlations of 3D ultrasound features and prognostic factors were analysed.

RESULTS: The retraction pattern in the coronal plane had a significant value as an independent predictor of a small tumour size (P #8201;= 0.014), a lower histological grade (P #8201;= 0.009) and positive ER or PR expression status (P #8201;= 0.001, 0.044).

The retraction pattern with a hyperechoic ring only existed in low-grade and ER-positive tumours.

The presence of the hyperechoic ring strengthened the ability of the retraction pattern to predict a good prognosis of breast cancer.

The increased intra-tumour vascularization index (VI, the mean tumour vascularity) reflected a higher histological grade (P #8201;= 0.025) and had a positive correlation with MVD (r #8201;= 0.530, P #8201;= 0.001).

CONCLUSIONS: The retraction pattern and histogram indices of VI provided by 3D ultrasound may be useful in predicting prognostic information about breast cancer.

KEY POINTS: • Three-dimensional ultrasound can potentially provide prognostic evaluation of breast cancer. • The retraction pattern and hyperechoic ring in the coronal plane suggest good prognosis. • The increased intra-tumour vascularization index reflects a higher histological grade. • The intra-tumour vascularization index is positively correlated with microvessel density.

Built by Lextutor, SD function added 18 Nov 2010

Part-of-speech tagging

- Assigning a syntatic tag to each word in a sentence

Stanford Parser

Please enter a sentence to be parsed:

Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included.

Language: English ▾ Sample Sentence Parse

Your query

Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included.

Tagging

```
Surgical/NNP resection/NN specimens/NNS of/IN 85/CD invasive/JJ
ductal/JJ carcinomas/NNS of/IN 85/CD women/NNS who/WP had/VBD
undergone/VBN 3D/CD ultrasound/NN were/VBD included/VBN ./.
```

Parsing

- Building the synthatic tree of a sentence

Parse

```
(ROOT
  (S
    (NP
      (NP (NNP Surgical) (NN resection) (NNS specimens))
      (PP (IN of)
        (NP
          (NP (CD 85) (JJ invasive) (JJ ductal) (NNS carcinomas))
          (PP (IN of)
            (NP
              (NP (CD 85) (NNS women))
              (SBAR
                (WHNP (WP who))
                (S
                  (VP (VBD had)
                    (VP (VBN undergone)
                      (NP (CD 3D) (NN ultrasound)))))))))))
    (VP (VBD were)
      (VP (VBN included)))
    (. .)))
```

Parsing

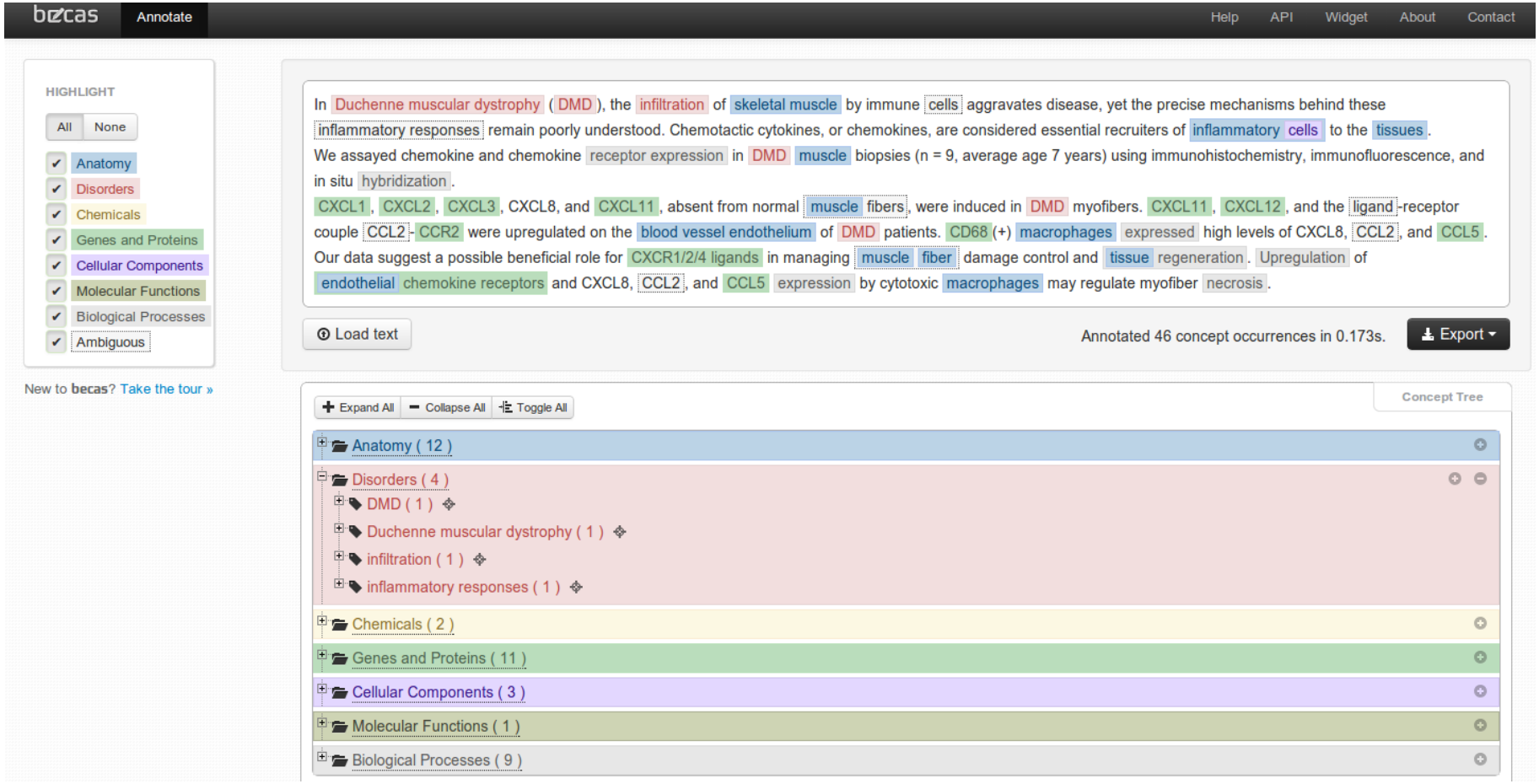
- Building the synthatic tree of a sentence

Typed dependencies

```
nn(specimens-3, Surgical-1)
nn(specimens-3, resection-2)
nsubjpass(included-18, specimens-3)
prep(specimens-3, of-4)
num(carcinomas-8, 85-5)
amod(carcinomas-8, invasive-6)
amod(carcinomas-8, ductal-7)
pobj(of-4, carcinomas-8)
prep(carcinomas-8, of-9)
num(women-11, 85-10)
pobj(of-9, women-11)
nsubj(undergone-14, who-12)
aux(undergone-14, had-13)
rcmod(women-11, undergone-14)
num(ultrasound-16, 3D-15)
dobj(undergone-14, ultrasound-16)
auxpass(included-18, were-17)
root(ROOT-0, included-18)
```

Named-entity recognition

- Identifying pre-defined entity types in a sentence



Word sense disambiguation

- Figuring out the exact meaning of a word or entity

Noun 1. tie - neckwear consisting of a long narrow piece of material worn (mostly by men) under a collar and tied in knot at the front; "he stood in front of the mirror tightening his necktie"; "he wore a vest and tie"

- necktie
- bola, bola tie, bolo, bolo tie - a cord fastened around the neck with an ornamental clasp and worn as a necktie
- bow tie, bow-tie, bowtie - a man's tie that ties in a bow
- four-in-hand - a long necktie that is tied in a slipknot with one end hanging in front of the other
- neckwear - articles of clothing worn about the neck
- old school tie - necktie indicating the school the wearer attended
- string tie - a very narrow necktie usually tied in a bow
- Windsor tie - a wide necktie worn in a loose bow

2. tie - a social or business relationship; "a valuable financial affiliation"; "he was sorry he had to sever his ties with other members of the team"; "many close associations with England"

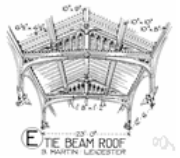
- affiliation, tie-up, association
- relationship - a state involving mutual dealings between people or parties or countries

3. tie - equality of score in a contest

- equivalence, par, equality, equation - a state of being essentially equal or equivalent; equally balanced; "on a par with the best"
- deuce - a tie in tennis or table tennis that requires winning two successive points to win the game

4. tie - a horizontal beam used to prevent two other structural members from spreading apart or separating; "he nailed the rafters together with a tie beam"

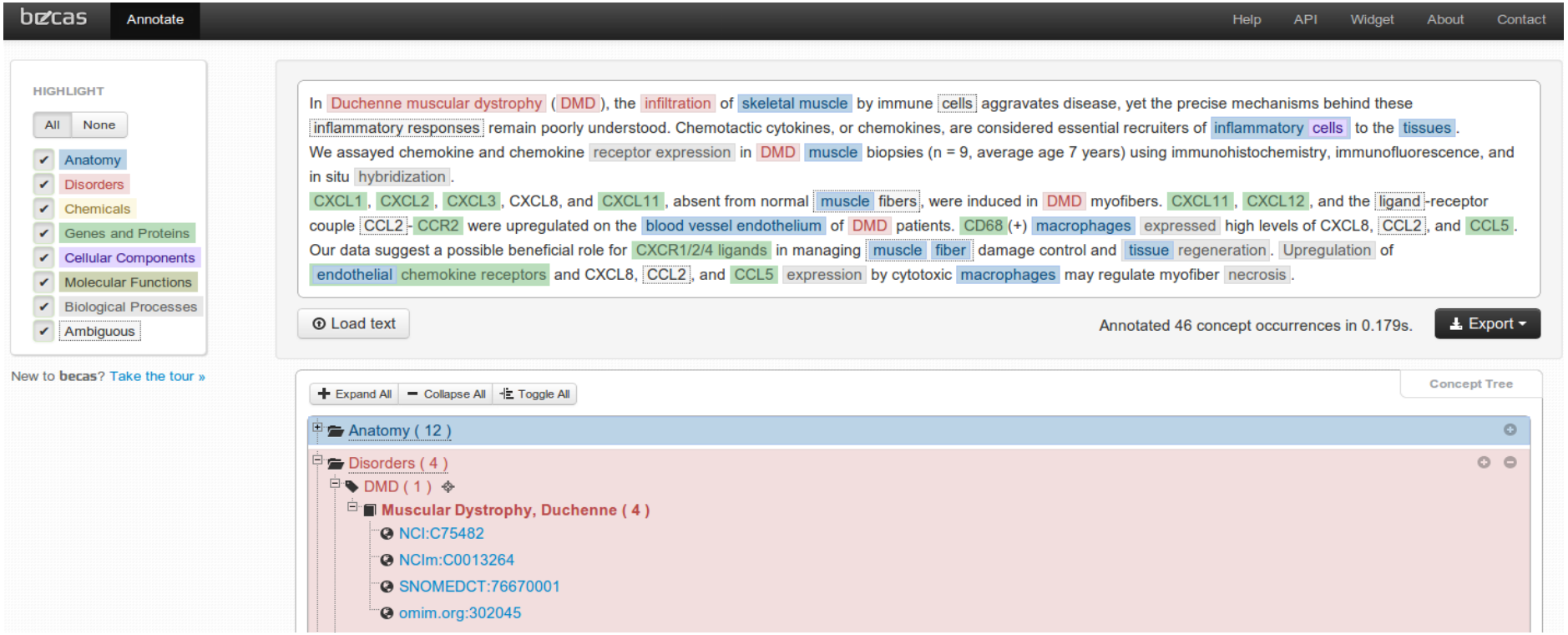
- tie beam
- beam - long thick piece of wood or metal or concrete, etc., used in construction



http://www.thefreedictionary.com/tie

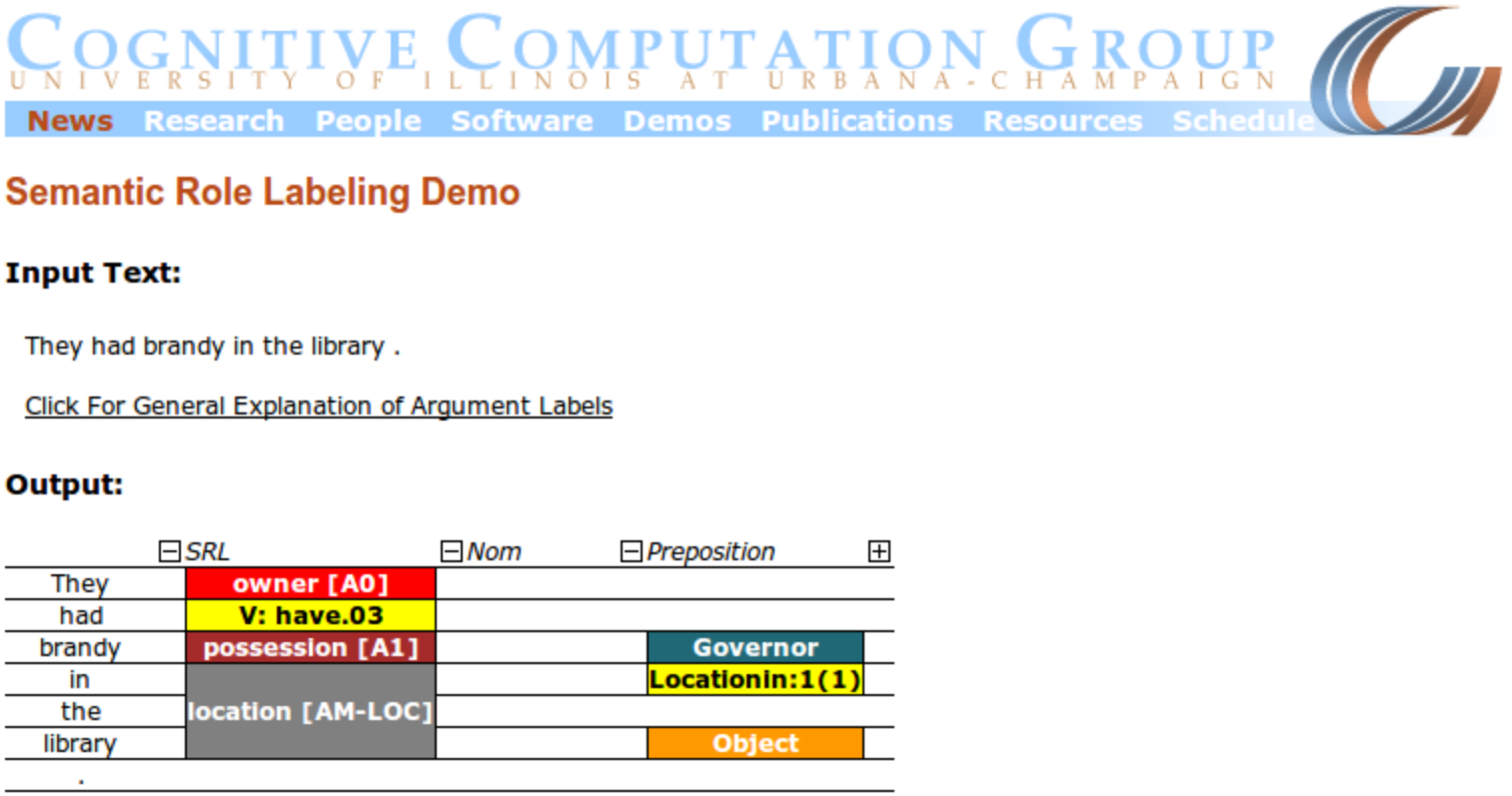
Word sense disambiguation

- Figuring out the exact meaning of a word or entity



Semantic role labeling

- Extracting subject-predicate-object triples from a sentence



Outline

- NLP course
- Introduction to NLP
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- Course materials

Phonetics and phonology

- The study of linguistic sounds and their relations to words

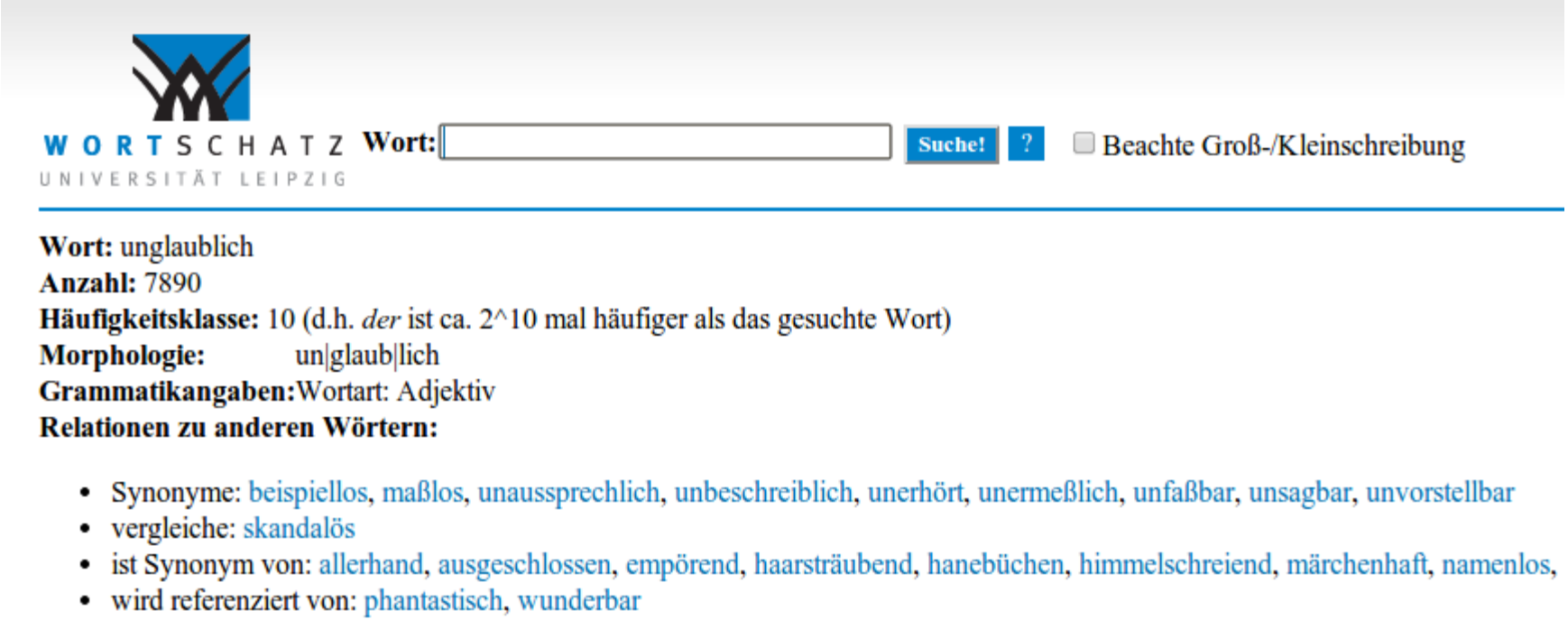
Das Funkalphabet - German Phonetic Spelling Code
compared to the international ICAO/NATO code
Listen to AUDIO for this chart! (below)

Germany*	Phonetic Guide	ICAO/NATO**
A wie **Anton**	AHN-tone	**Alfa/Alpha**
Ä wie **Ärger**	AIR-gehr	(1)
B wie **Berta**	BARE-tuh	**Bravo**
C wie **Cäsar**	SAY-zar	**Charlie**
Ch wie **Charlotte**	shar-LOT-tuh	(1)
D wie **Dora**	DORE-uh	**Delta**
E wie **Emil**	ay-MEAL	**Echo**
F wie **Friedrich**	FREED-reech	**Foxtrot**
G wie **Gustav**	GOOS-tahf	**Golf**
H wie **Heinrich**	HINE-reech	**Hotel**
I wie **Ida**	EED-uh	**India/Indigo**
J wie **Julius**	YUL-ee-oos	**Juliet**
K wie **Kaufmann**	KOWF-mann	**Kilo**
L wie **Ludwig**	LOOD-vig	**Lima**
AUDIO 1 > Listen to mp3 for A-L		
M wie **Martha**	MAR-tuh	**Mike**
N wie **Nordpol**	NORT-pole	**November**
O wie **Otto**	AHT-toe	**Oscar**
Ö wie **Ökonom** (2)	UEH-ko-nome	(1)
P wie **Paula**	POW-luh	**Papa**
Q wie **Quelle**	KVEL-uh	**Quebec**
R wie **Richard**	REE-shart	**Romeo**
S wie **Siegfried** (3)	SEEG-freed	**Sierra**
Sch wie **Schule**	SHOO-luh	(1)
ß (**Eszett**)	ES-TSET	(1)
T wie **Theodor**	TAY-oh-dore	**Tango**
U wie **Ulrich**	OOL-reech	**Uniform**
Ü wie **Übermut**	UEH-ber-moot	(1)
V wie **Viktor**	VICK-tor	**Victor**
W wie **Wilhelm**	VIL-helm	**Whiskey**
X wie **Xanthippe**	KSAN-tipp-uh	**X-Ray**
Y wie **Ypsilon**	IPP-see-lohn	**Yankee**
Z wie **Zeppelin**	TSEP-puh-leen	**Zulu**

http://german.about.com/library/blfunkabc.htm

Morphology

- The study of internal structures of words and how they can be modified
- Parsing complex words into their components



WORTSCHATZ UNIVERSITÄT LEIPZIG **Wort:** Suche! ? Beachte Groß-/Kleinschreibung

Wort: unglaublich
Anzahl: 7890
Häufigkeitsklasse: 10 (d.h. *der* ist ca. 2^10 mal häufiger als das gesuchte Wort)
Morphologie: un|glaub|lich
Grammatikangaben: Wortart: Adjektiv
Relationen zu anderen Wörtern:

- Synonyme: beispiellos, maßlos, unaussprechlich, unbeschreiblich, unerhört, unermeßlich, unfaßbar, unsagbar, unvorstellbar
- vergleiche: skandalös
- ist Synonym von: allerhand, ausgeschlossen, empörend, haarsträubend, hanebüchen, himmelschreiend, märchenhaft, namenlos,
- wird referenziert von: phantastisch, wunderbar

Syntax

- The study of the structural relationships between words in a sentence

Parse

```
(ROOT
  (S
    (NP
      (NP (NNP Surgical) (NN resection) (NNS specimens))
      (PP (IN of)
        (NP
          (NP (CD 85) (JJ invasive) (JJ ductal) (NNS carcinomas))
          (PP (IN of)
            (NP
              (NP (CD 85) (NNS women))
              (SBAR
                (WHNP (WP who))
                (S
                  (VP (VBD had)
                    (VP (VBN undergone)
                      (NP (CD 3D) (NN ultrasound)))))))))))
    (VP (VBD were)
      (VP (VBN included)))
    (. .)))
```

Semantics

- The study of the meaning of words, and how these combine to form the meanings of sentences
  - Synonymy: fall & autumn
  - Hypernymy, hyponymy (is a): dog & animal
  - Meronymy (part of): finger & hand
  - Homonymy: fall (verb & season)
  - Antonymy: big & small

Pragmatics

- Social language use
- The study of how language is used to accomplish goals, and the influence of context on meaning
- Understanding the aspects of a language which depends on situation and world knowledge

Give me the salt!

Could you please give me the salt?

Discourse

- The study of linguistic units larger than a single statement

John reads a book. He borrowed it from his friend.

Outline

- NLP course
- Introduction to NLP
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- Course materials

Paraphrasing

- Different words/sentences express the same meaning
  - Season of the year
    - Fall
    - Autumn
  - Book delivery time
    - When will my book arrive?
    - When will I receive my book?

Ambiguity

- One word/sentence can have different meanings
  - Fall
    - The third season of the year
    - Moving down towards the ground or towards a lower position
  - The door is open.
    - Expressing a fact
    - A request to close the door

Phonetics and Phonology

Communication tip:

Phonological ambiguities or **Give peas a chance!**

One of my favourite ways to have fun with communication are phonological ambiguities.
Phonological ambiguities are two or more words which sound the same and have different meanings.

Language can contain ambiguities - and more than one way to compose a set of sounds into words.

So listen to yourself: It is always good to notice a spoken sentence often contains many words which are (sometimes not) intended to be heard.

English examples:

- there – their
- here - hear
- plane – plain
- Hamburger (Citizens of Hamburg) – hamburger (burger, food)
- sea - see
- Friday - fry day
- weekend - weak end
- ice cream - I scream.
- new direction - nude erection
- new day - nude, eh?
- I don't know! - I don't - no!
- but – butt
- Wait - Weight
- psychotherapist - psycho the rapist
- You're unconscious now... - Your unconscious now...
- Your students... - You're students...
- Two - too - to

German examples:

- Du hast Gewehre. (You have got guns.) - Du hasst Gewehre. (You hate guns.)
- Lehrer (teacher) – leerer (emptier)

http://worldsgreatestsmile.com/html/phonological_ambiguity.html

Syntax and ambiguity

- I saw the man with a telescope.
  - Who had the telescope?

Semantics

- The astronomer loves the star.
  - Star in the sky
  - Celebrity

Discourse

- Alice understands that you like your mother, but she …
  - Does she refer to Alice or your mother?

Outline

- NLP course
- Introduction to NLP
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- Course materials

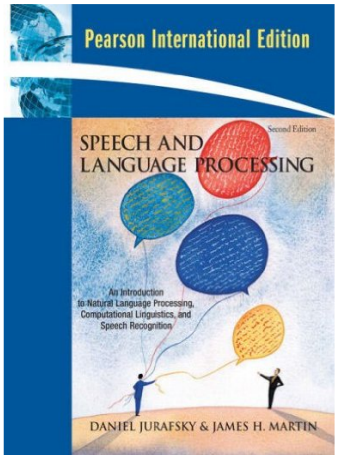
Topics

Week	Date	Topic
1	April 9th, 2014	(no lecture)
2	April 16th, 2014	Introduction to Language Technology
3	April 23rd, 2014	Language Modeling
4	April 30th, 2014	Machine Learning for NLP
5	May 7th, 2014	Part Of Speech Tagging, Syntactic Analysis
6	May 14th, 2014	Named Entity Recognition, Word Similarity, Word Sense Disambiguation
7	May 21st, 2014	Lexical Semantics, Semantic Role Labeling
8	May 28th, 2014	(no lecture)
9	June 4th, 2014	(no lecture)
10	June 11th, 2014	Information Retrieval, Text Categorization
11	June 18th, 2014	Information Extraction, Ontology Extraction
12	June 25th, 2014	Summarization, Question Answering
13	July 2nd, 2014	Sentiment Analysis, Machine Translation
14	July 9th, 2014	Review, Q&A
15	July 16th, 2014	Exam

Exercise	Due	Topic
1	June 11th, 2014	to be decided
2	July 9th, 2014	to be decided

Course book

- Speech and Language Processing
  - Daniel Jurafsky and James H. Martin



Speech and Language Processing von Daniel Jurafsky und James H. Martin von Prentice Hall International (17. Juli 2013)

EUR 56,51 Taschenbuch *Prime*

Bestellen Sie in den nächsten **5 Stunden**, um den Artikel am **Dienstag, 15. April** zu erhalten.

Nur noch 1 Stück auf Lager - jetzt bestellen.

EUR 39,56 Kindle Edition

Jetzt als Download verfügbar.

Andere Angebote - Taschenbuch

EUR 53,27 neu (34 Angebote)

EUR 48,99 gebraucht (2 Angebote)

Bestseller Nr. 1 in Software zur Verarbeitung natürlicher Sprache

Kostenlose Lieferung möglich.

Standort:	Handapparat
Ausleihstatus:	eingeschraenkte Benutzung Bestellen und Vormerken ueber den OPAC ist nicht moeglich.
Standort:	Bereichsbibliothek Babelsberg --> Wegweiser
Signatur:	**ST 306 JUR**
Ausleihstatus:	Ausleihbar - LBS ausgeliehen bis 09-05-2014 ▸ Vormerken
Standort:	Bereichsbibliothek Babelsberg --> Wegweiser
Signatur:	**ST 306 JUR**
Ausleihstatus:	Ausleihbar - LBS Verfuegbar: BB Babelsberg / LBS.
Standort:	Bereichsbibliothek Babelsberg --> Wegweiser
Signatur:	**ST 306 JUR**
Ausleihstatus:	Ausleihbar ausgeliehen bis 08-05-2014 ▸ Vormerken
Standort:	Bereichsbibliothek Babelsberg --> Wegweiser
Signatur:	**ST 306 JUR**
Ausleihstatus:	Ausleihbar - LBS Verfuegbar: BB Babelsberg / LBS.
Standort:	Bereichsbibliothek Babelsberg --> Wegweiser
Signatur:	**ST 306 JUR**
Ausleihstatus:	Ausleihbar - LBS Verfuegbar: BB Babelsberg / LBS.
Standort:	Bereichsbibliothek Babelsberg --> Wegweiser
Signatur:	**ST 306 JUR**
Ausleihstatus:	Ausleihbar - LBS Verfuegbar: BB Babelsberg / LBS.
Standort:	Bereichsbibliothek Babelsberg --> Wegweiser
Signatur:	**ST 306 JUR**
Ausleihstatus:	Praesenzbestand Verfuegbar: BB Babelsberg / Präsenz.

Journal and conferences

- Journal
  - Computational Linguistics
- Conferences
  - ACL: Association for Computational Linguistics
  - NAACL: North American Chapter
  - EACL: European Chapter
  - HLT: Human Language Technology
  - EMNLP: Empirical Methods on Natural Language Processing
  - CoLing: Computational Linguistics
  - LREC: Language Resources and Evaluation