Natural Language Processing

SoSe 2017

HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Introduction to Natural Language Processing

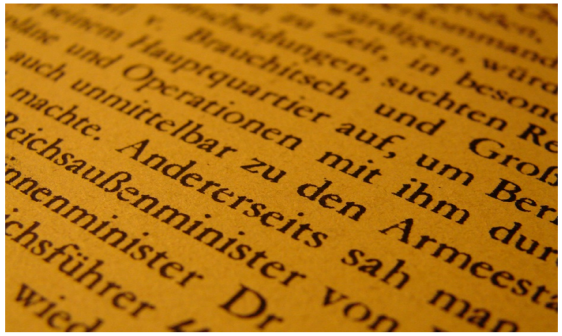
Dr. Mariana Neves

April 24th, 2017

Outline

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

Natural Language



日本語で

冬は世界各地でさまざまなお祝いが行われる時期です。ほんのいくつか例を挙げるだけでも、ハナカ、クリスマス、クワンザ、新年などさまざまなお祝いがあります。各文化によってその祝い方はさまざまですが、ほとんどのお祝いにはごちそうが欠かせません。

(http://expertenough.com/2392/german-language-hacks)

(http://www.transparent.com/learn-japanese/articles/dec_99.html)

Artificial Language

```
try {
    cMessage = messageQueue.take();
    for (AsyncContext ac : queue) {
        try {
            PrintWriter acWriter = ac.g
            acWriter.println(cMessage);
            acWriter.flush();
        } catch(IO
            System
            queue.
        }
    }
} catch(Interrupte
```

append(CharSequenc
append(char c)
append(CharSequenc
format(String forma
format(Locale l, S
printf(String forma

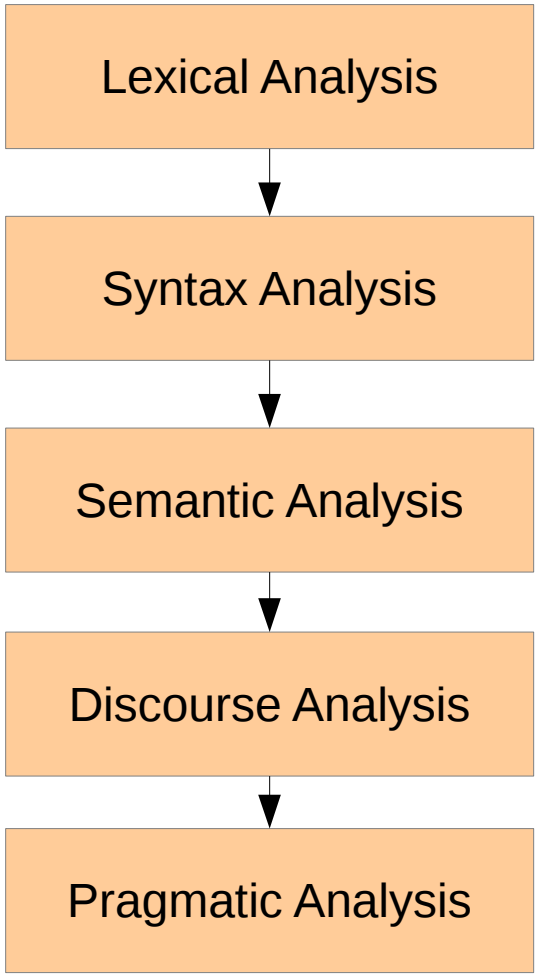
```
def add5(x):
    return x+5

def dotwrite(ast):
    nodename = getNodename()
    label=symbol.sym_name.get(int(ast[0]),ast[0])
    print '    %s [label="%s' % (nodename,label),
    if isinstance(ast[1], str):
        if ast[1].strip():
            print '= %s"];' % ast[1]
        else:
            print '"]'
    else:
        print '"];'
        children = []
        for n, child in enumerate(ast[1:]):
            children.append(dotwrite(child))
        print '    %s -> {' % nodename,
        for name in children:
            print '%s' % name,
```

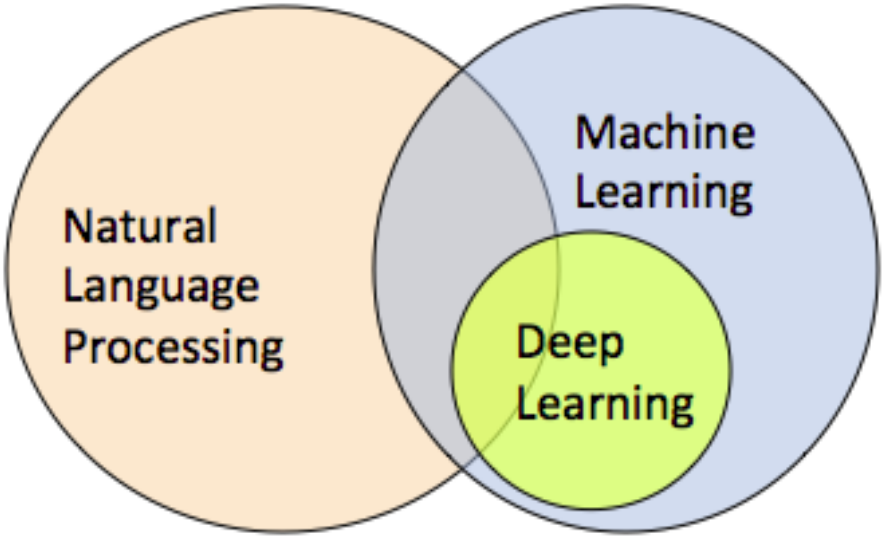
(http://noobite.com/learn-programming-start-with-python/)

(https://netbeans.org/features/java/)

Natural Language Processing



Natural Language Processing



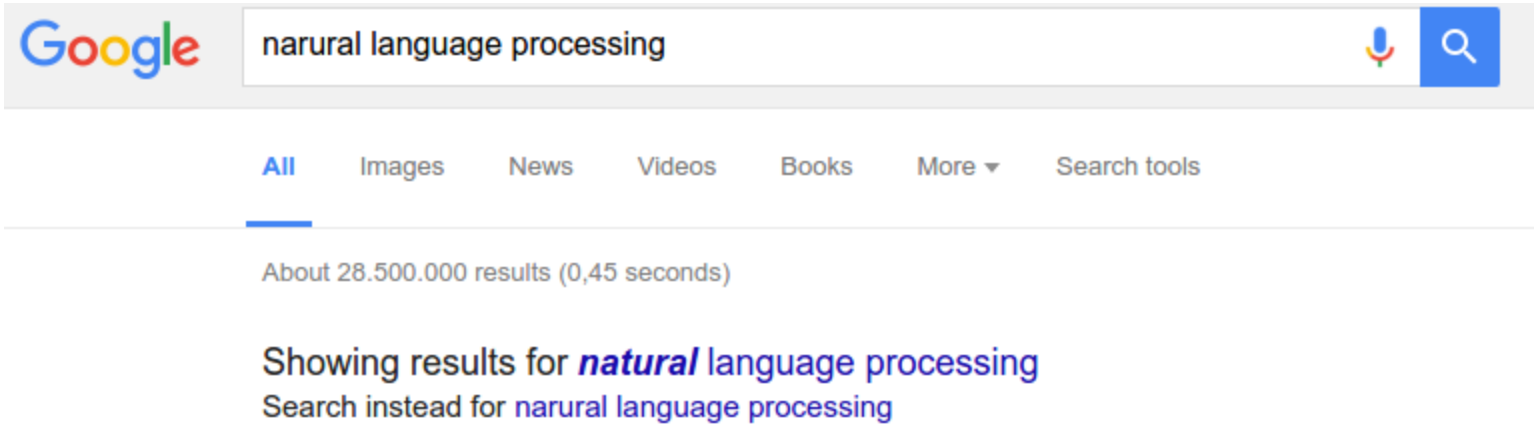
(http://rutumulkar.com/blog/2016/NLP-ML)

Outline

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

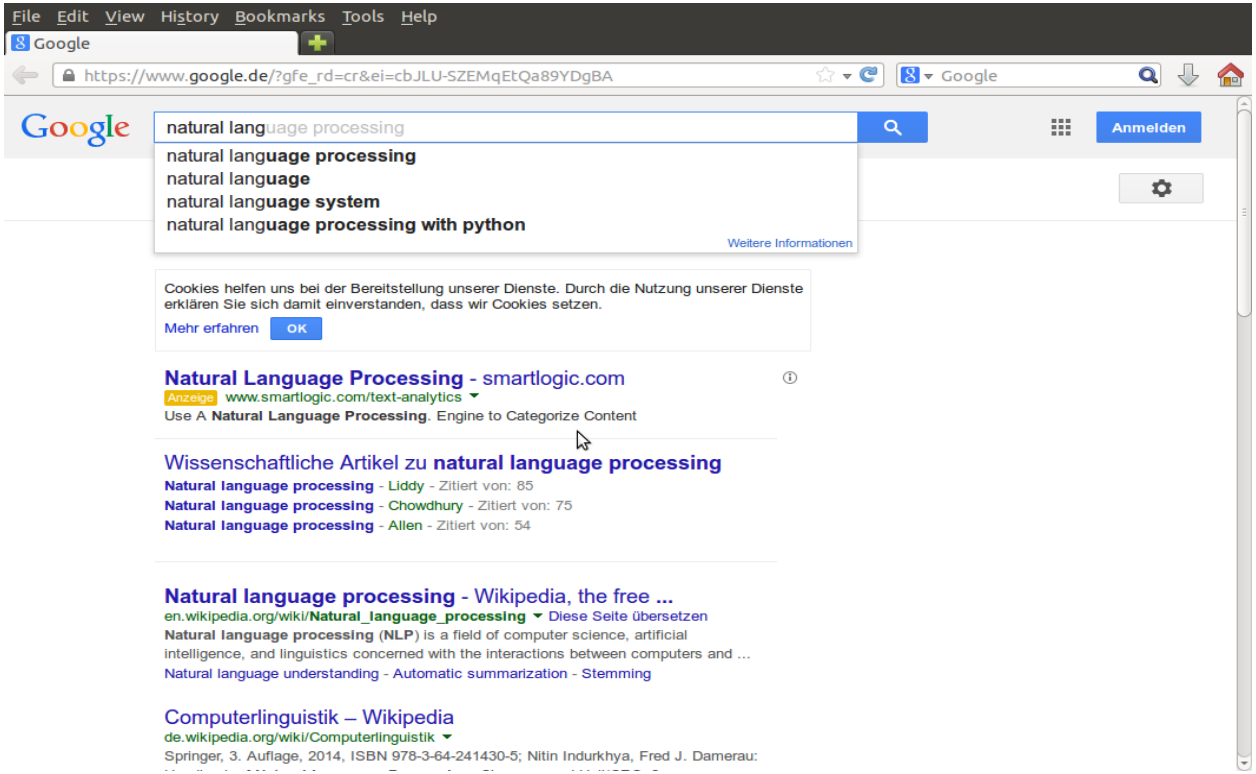
Spell and Grammar Checking

- Checking spelling and grammar
- Suggesting alternatives for the errors



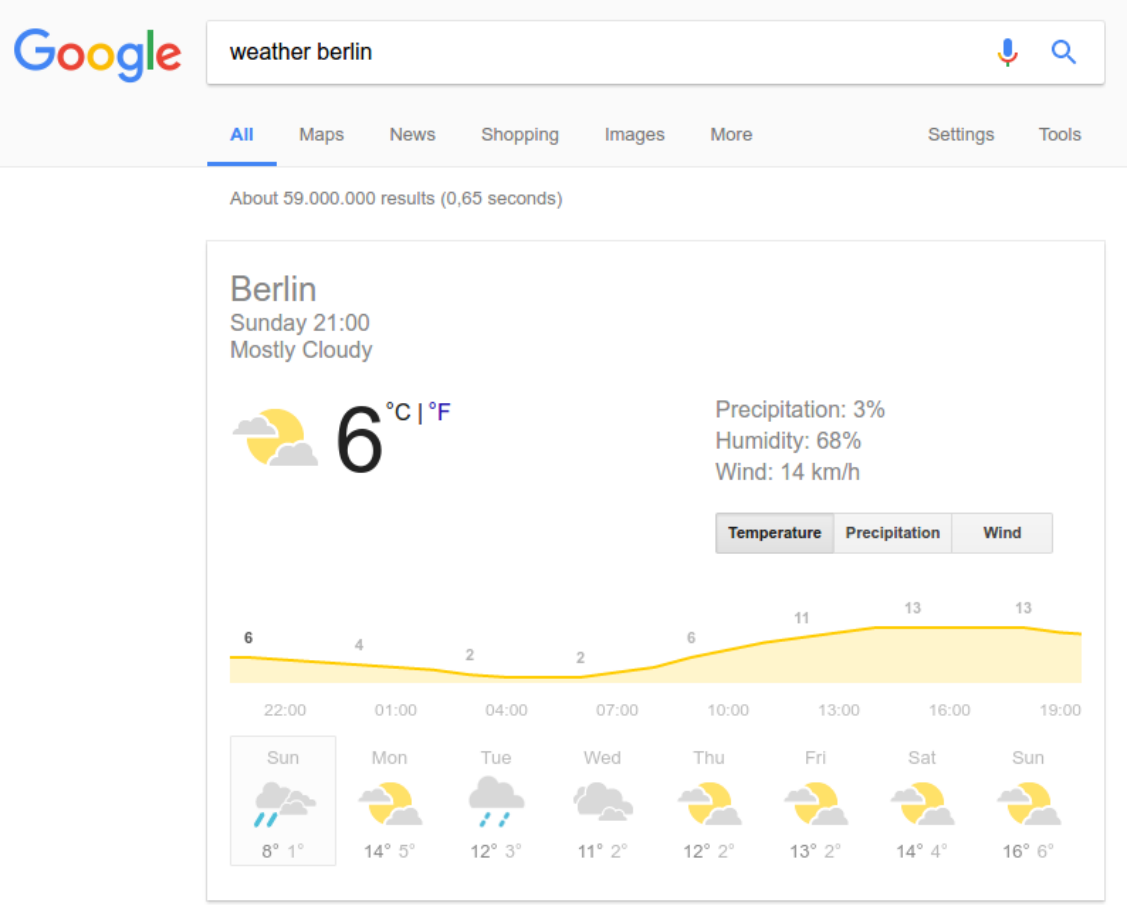
Word Prediction

- Predicting the most probable next word to be typed by the user



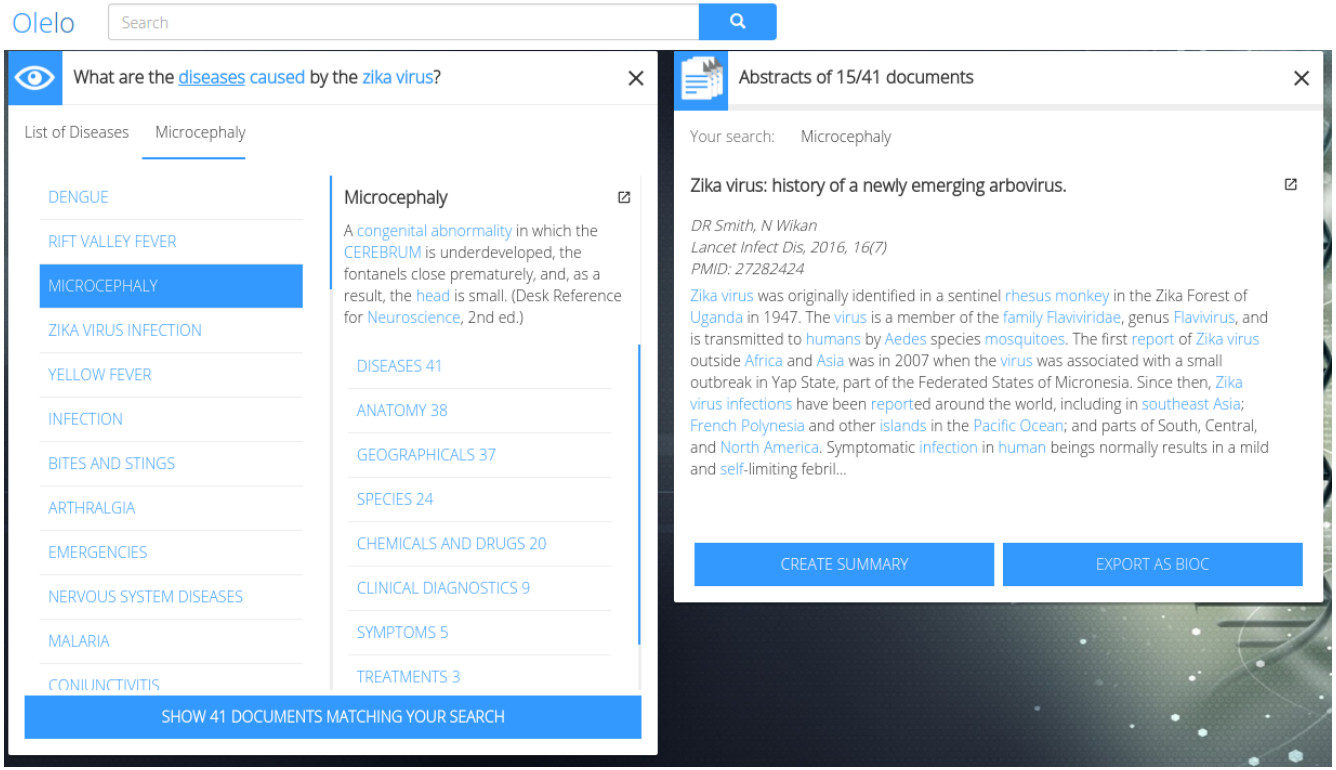
Information Retrieval

- Finding relevant information to the user's query



Information Retrieval

- Finding relevant information to the user's query



(https://hpi.de/plattner/olelo)

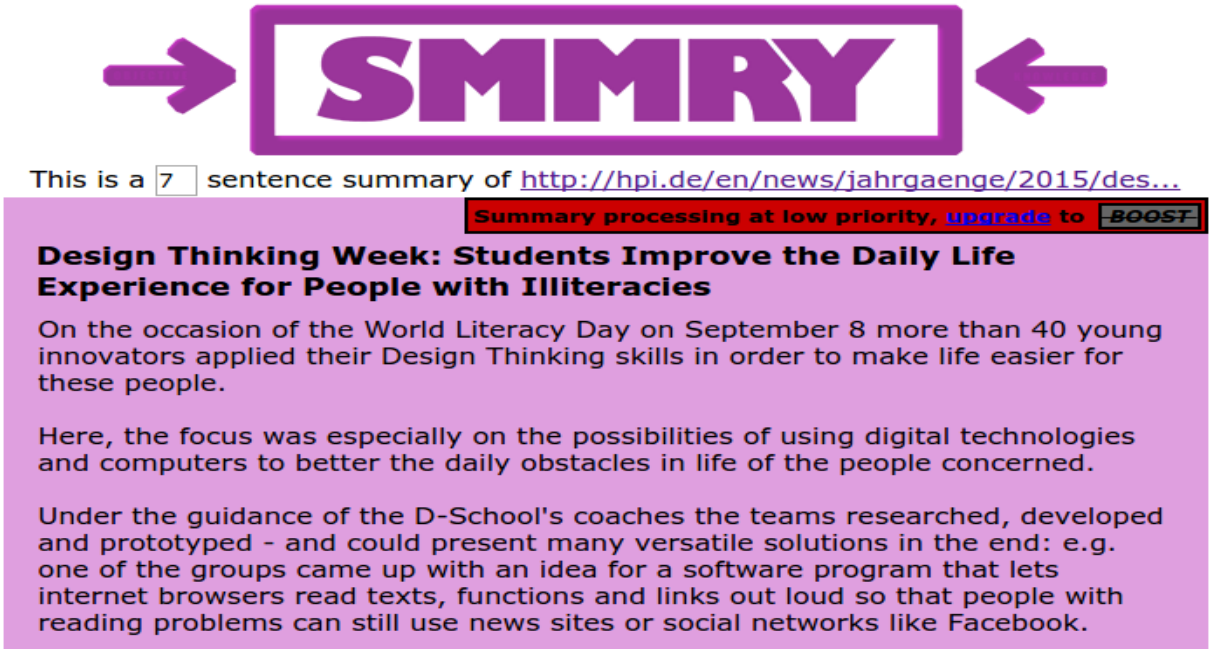
Text Categorization

- Assigning one (or more) pre-defined category to a text

http://www.uclassify.com/browse/mvazquez/News-Classifier

Summarization

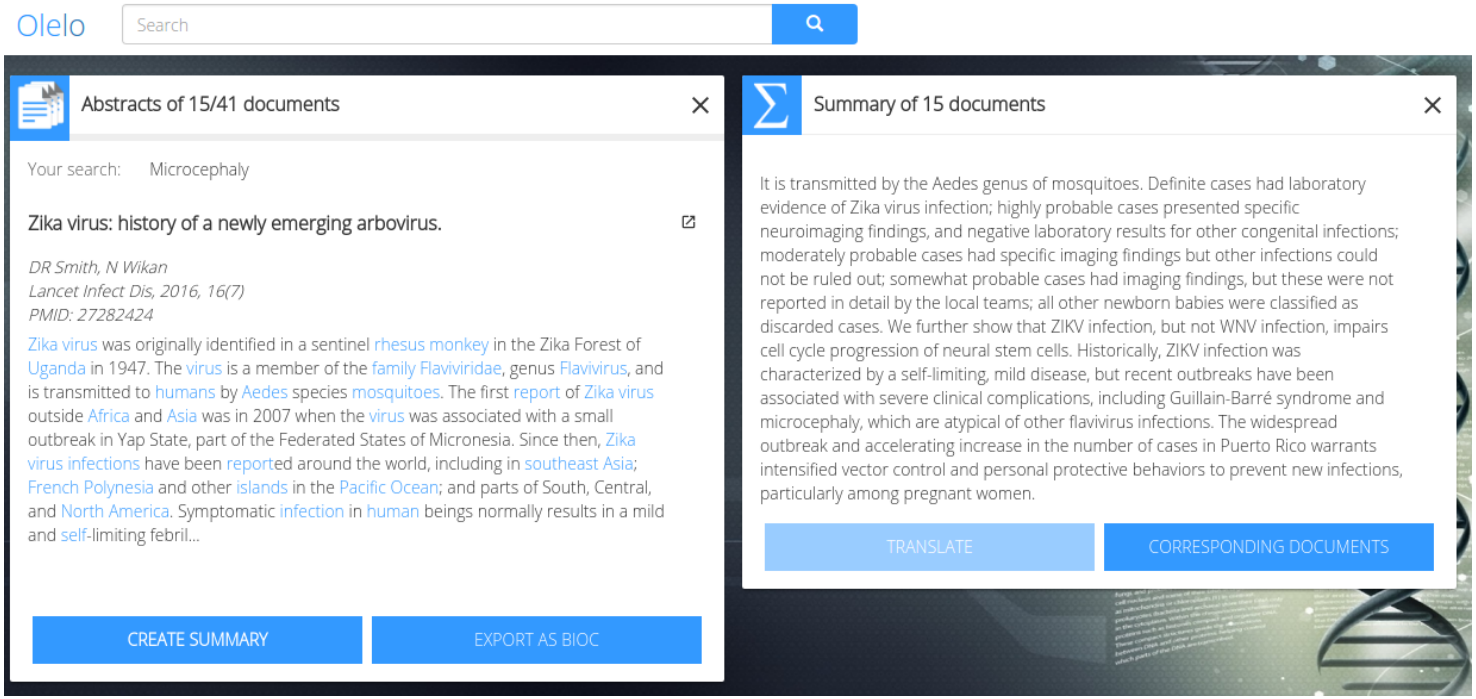
- Generating a short summary from one or more documents, sometimes based on a given query



http://smmry.com/

Summarization

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



(https://hpi.de/plattner/olelo)

Question answering

- Answering questions with a short answer

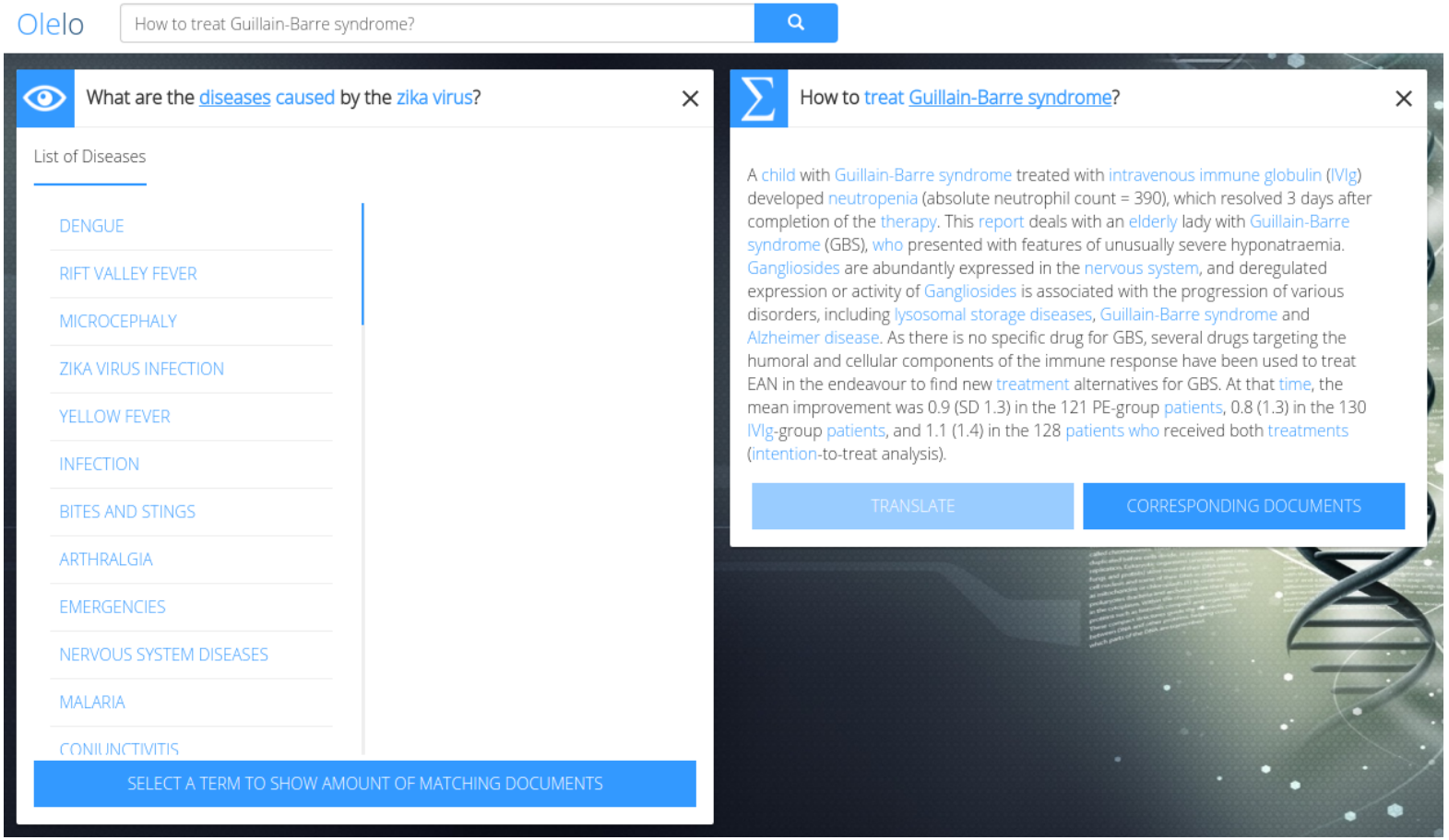
===> what countries speak Spanish

The language Spanish is spoken in Argentina, Aruba, Belize, Bolivia, Brazil, Canada, Cayman Islands, Chile, Colombia, Costa Rica, Cuba, Curacao, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Falkland Islands (Islas Malvinas), Gibraltar, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Saint Martin, Sint Maarten, Spain, Switzerland, Trinidad and Tobago, United States, Uruguay, Venezuela, and Virgin Islands.

The language Castilian Spanish is spoken in Spain.

http://start.csail.mit.edu/index.php

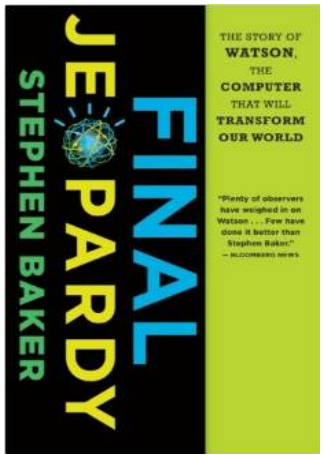
Question Answering



(https://hpi.de/plattner/olelo)

Question answering

- IBM Watson in Jeopardy



https://www.youtube.com/watch?v=WFR3lOm_xhE

Information Extraction

- Extracting important concepts from texts and assigning them to slot in a certain template

Angela Merkel

Merkel at the EPP Summit, March 2016

Chancellor of Germany	
Incumbent	
Assumed office	22 November 2005
President	Horst Köhler Christian Wulff Joachim Gauck
Deputy	Franz Müntefering Frank-Walter Steinmeier Guido Westerwelle Philipp Rösler Sigmar Gabriel
Preceded by	Gerhard Schröder
Leader of the Christian Democratic Union	
Incumbent	
Assumed office	10 April 2000
Preceded by	Wolfgang Schäuble
Minister for the Environment	
In office	17 November 1994 – 26 October 1998
Chancellor	Helmut Kohl
Preceded by	Klaus Töpfer
Succeeded by	Jürgen Trittin
Minister for Women and Youth	
In office	18 January 1991 – 17 November 1994
Chancellor	Helmut Kohl
Preceded by	Ursula Lehr
Succeeded by	Claudia Nolte
Personal details	
Born	Angela Dorothea Kasner 17 July 1954 (age 61) Hamburg, West Germany
Political party	Democratic Awakening (1989–1990) Christian Democratic Union (1990–present)
Spouse(s)	Ulrich Merkel (1977–1982) Joachim Sauer (1998–present)
Alma mater	Leipzig University
Religion	Lutheranism (within Evangelical Church)
Signature	

Information Extraction

- Includes named-entity recognition

Angela Merkel

From Wikipedia, the free encyclopedia

"Merkel" redirects here. For other uses, see Merkel (disambiguation).

Angela Dorothea Merkel (English /ˈæŋgələ ˈmɛərkəl/[a]; née **Kasner**; born 17 July 1954) is a German politician who is currently Chancellor of Germany. She is also the leader of the Christian Democratic Union (CDU). Merkel has been described at various times as the *de facto* leader of the European Union, the most powerful woman in the world, and the world's second most powerful person.

A former research scientist with a doctorate in physical chemistry, Merkel entered politics in the wake of the Revolutions of 1989, and briefly served as a deputy spokesperson for the first democratically elected East German Government headed by Lothar de Maizière in 1990. Following German reunification in 1990, Merkel was elected to the Bundestag for the state of Mecklenburg-Vorpommern, and has been reelected ever since. Merkel was appointed as the Minister for Women and Youth in the federal government under Chancellor Helmut Kohl in 1991, and became the Minister for the Environment in 1994. After her party lost the federal election in 1998, Merkel was elected Secretary-General of the CDU before becoming the party's first woman leader two years later in the aftermath of a donations scandal that toppled Wolfgang Schäuble.

Following the 2005 federal election, Merkel was appointed Germany's first woman Chancellor at the head of a grand coalition consisting of the CDU, its Bavarian sister party, the Christian Social Union (CSU), and the Social Democratic Party of Germany (SPD). In the 2009 federal election, the CDU obtained the largest share of the vote and Merkel was able to form a coalition government with the support of the Free Democratic Party (FDP).[9] At the 2013 federal election, Merkel's CDU won a landslide victory with 41.5% of the vote and formed a second grand coalition with the SPD, after the FDP lost all of its representation in the Bundestag.[10]

In 2007, Merkel was President of the European Council and chaired the G8, the second woman to do so. Merkel played a central role in the negotiation of the Treaty of Lisbon and the Berlin Declaration. One of Merkel's consistent priorities has been to strengthen transatlantic economic relations. Merkel played a crucial role in managing the financial crisis at the European and international level, and she has been referred to as "the decider." In domestic policy, health care reform, problems concerning future energy development and more recently her government's approach to the ongoing migrant crisis have been major issues during her Chancellorship.[11] On 26 March 2014, Merkel became the longest-serving incumbent head of government in the European Union and she is currently the senior G7 leader. On 20 November 2016, Merkel announced she would seek re-election to a fourth term.[12]

(https://en.wikipedia.org/wiki/Angela_Merkel)

Information Extraction

Gynecol Oncol. 2007 Dec;107(3):518-25. Epub 2007 Oct 25.

Combination therapy with pegylated liposomal doxorubicin and carboplatin in gynecologic malignancies: a prospective phase II study of the Arbeitsgemeinschaft Gynäekologische Onkologie Studiengruppe Ovarialkarzinom (AGO-OVAR) and Kommission Uterus (AGO-K-Ut).

du Bois A[1], Pfisterer J, Burchardi N, Loibl S, Huober J, Wi
Onkologie Studiengruppe Ovarialkarzinom; Kommission U

Author information

Abstract

OBJECTIVE: A multicenter non-randomized phase
doxorubicin (PLD) in combination with carboplatin

METHODS: One hundred forty women with recurr
sarcomas (n=11), or recurrent platinum-sensitive
every 28 days.

RESULTS: Hematological toxicities with NCI-CTC
febrile neutropenia in 2% of 652 cycles. Grade 3/
(9%), palmar-plantar erythrodysesthesia (7%), an
carboplatin. Seventy-four percent of all non-progr
evaluable for response): ovarian cancer (n=54) 6
cancer (n=26) 12%. Median progression-free surv
12.6) for endometrial cancer. Median overall survi

CONCLUSIONS: The combination of PLD and carl
Efficacy was low in cervical/vaginal cancer, but pr
recurrent platinum-sensitive ovarian cancer and is
comparing PLD/carboplatin with paclitaxel/carbop

PMID: 17910981 DOI: 10.1016/j.ygyno.2007.08.008

Evidence / Table

ID	**[[25345_dubois2007]]**			
Kapitel	Vaginalkarzinom			
	Endometriumkarzinom Ovarialkarzinom Zervixkarzinom			
Zuordnung	4_cht_pall_vagina 4_cht_pall_endo 4_cht_pall_ovar 4_cht_pall_zervix	PDF		
Typ	NCT			
Intervention(en)	Cth	Pall	Pegyliertes liposomales Doxorubicin +	40 mg/m$_2$ + AUC 6 Alle 4 Wochen

(https://www.ncbi.nlm.nih.gov/pubmed/17910981)

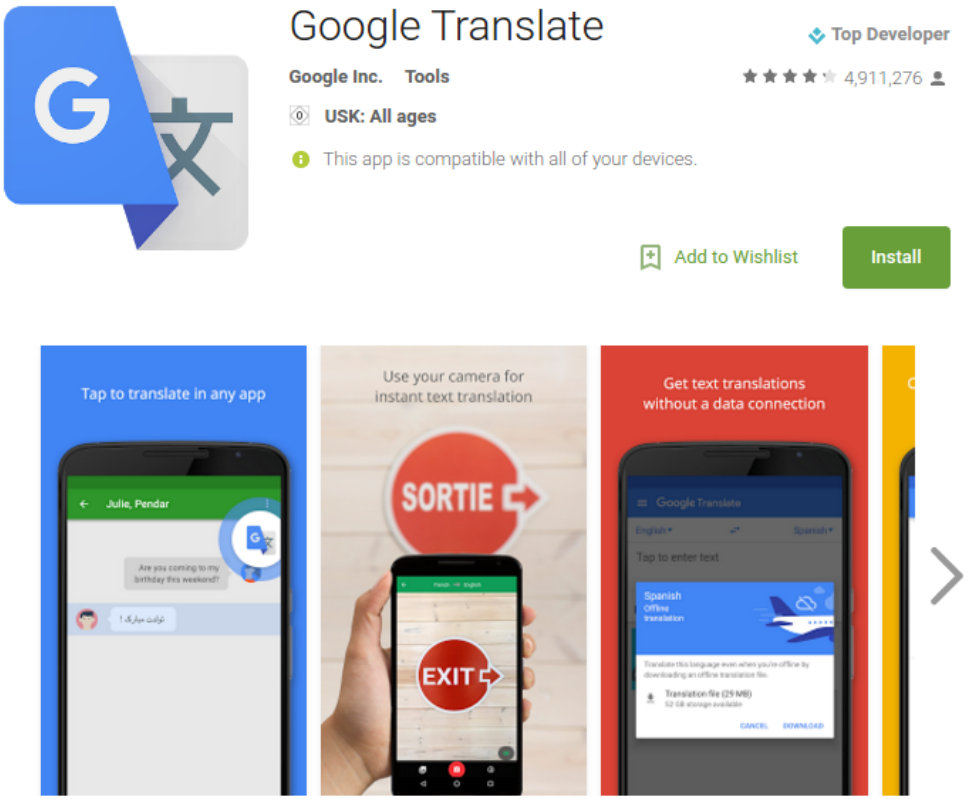
Machine Translation

- Translating a text from one language to another

(http://www.lemonde.fr/)

Machine Translation

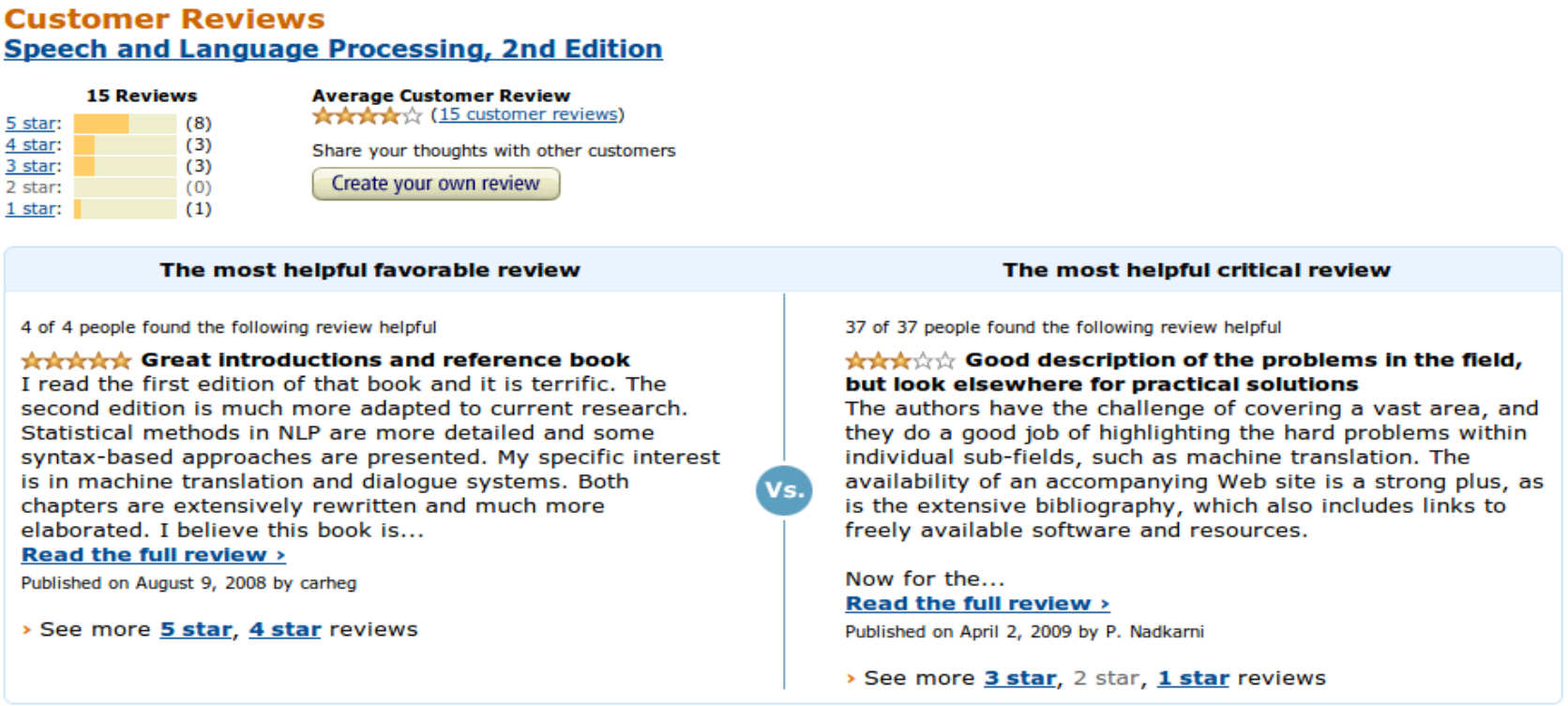
- Translating a text from one language to another



(https://www.youtube.com/watch?v=UtnXrXZYg_g)

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

Optical Character Recognition

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

Speech recognition

- Recognizing a spoken language and transforming it into a text

Siri.
Your wish is
its command.

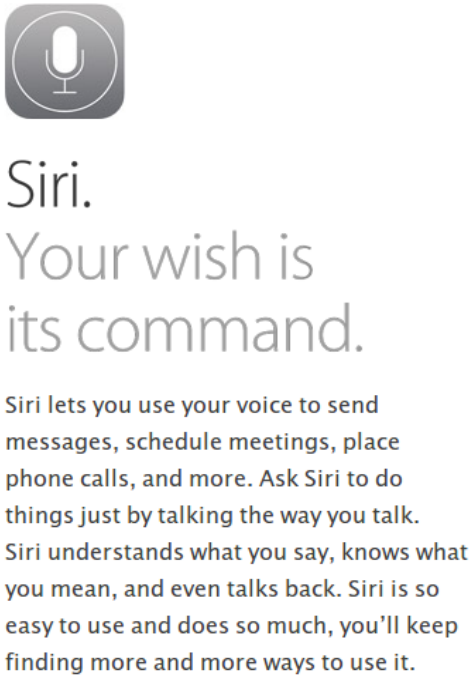
Siri lets you use your voice to send messages, schedule meetings, place phone calls, and more. Ask Siri to do things just by talking the way you talk. Siri understands what you say, knows what you mean, and even talks back. Siri is so easy to use and does so much, you'll keep finding more and more ways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

Speech synthesis

- Producing a spoken language from a text

Spoken dialog systems

- Running a dialog between the user and the system



Level of difficulties

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

Level of difficulties

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

Level of difficulties

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

Outline

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

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

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

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.

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).

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 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).

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).

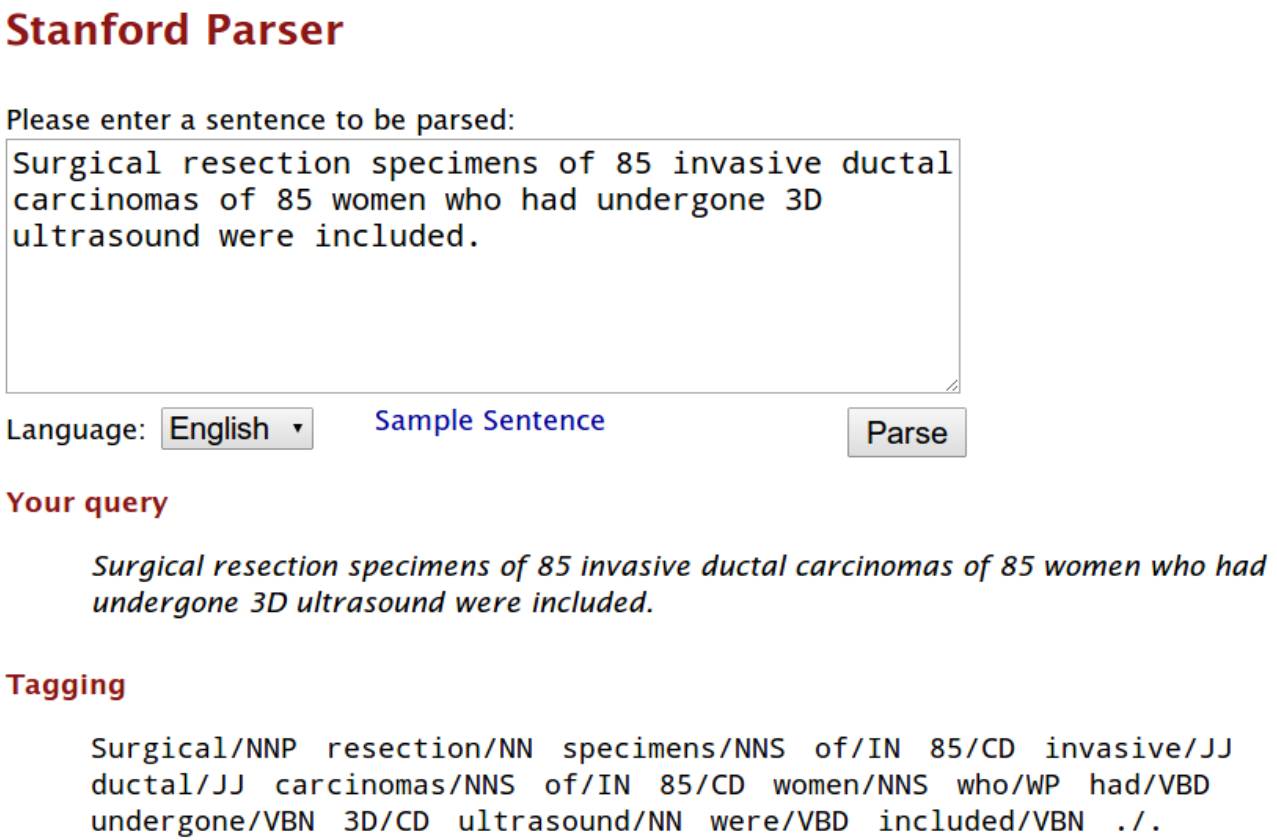
CONCLUSIONS: The retraction pattern and histogram indices of VI provided by 3D ultrasound may be useful in predicting prognostic information about breast cancer.

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



http://nlp.stanford.edu:8080/corenlp/

Parsing

- Building the syntactic 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)))
    (. .)))
```

http://nlp.stanford.edu:8080/corenlp/

Parsing

- Building the syntactic tree of a sentence

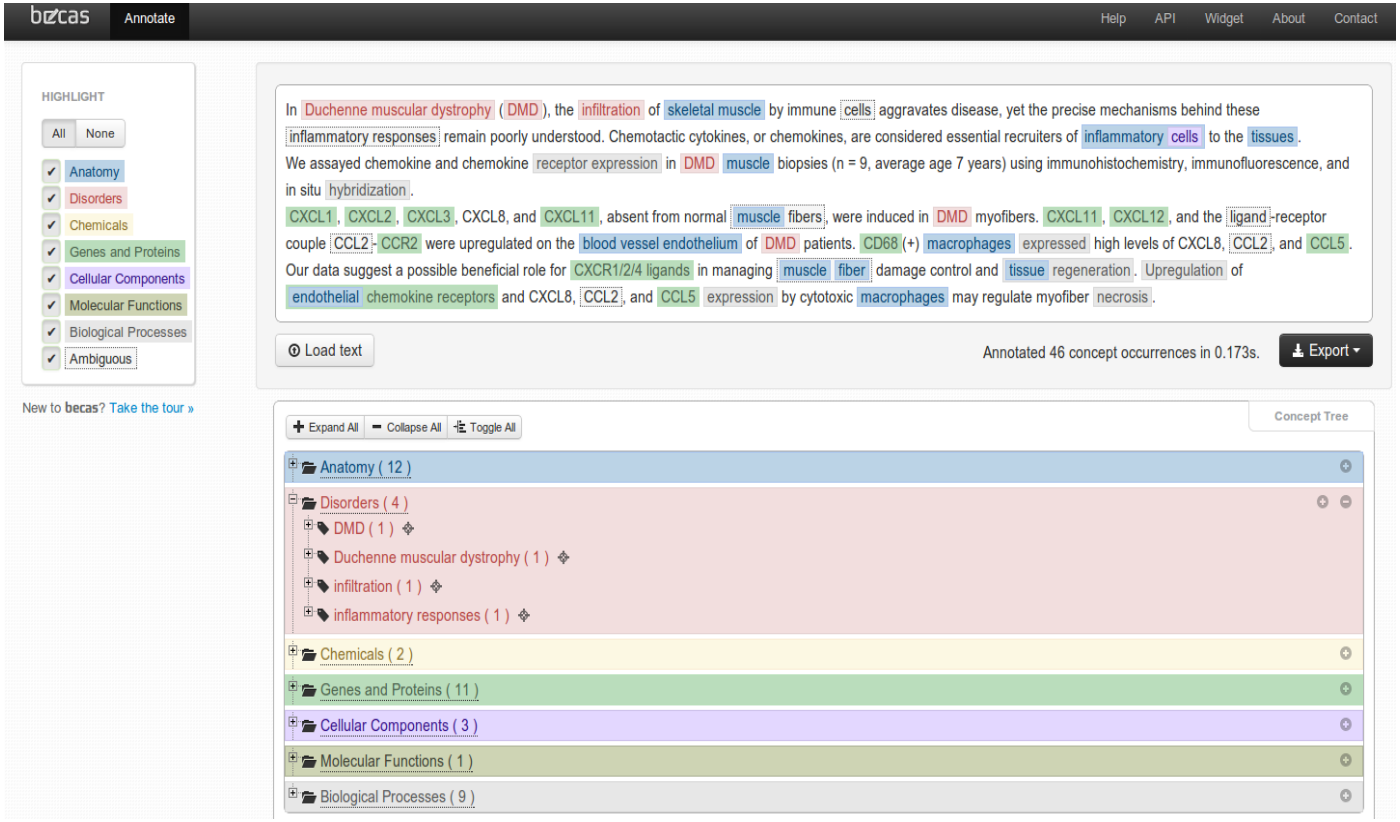
Typed dependencies

```
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)
```

http://nlp.stanford.edu:8080/corenlp/

Named-entity recognition

- Identifying pre-defined entity types in a sentence



Word sense disambiguation

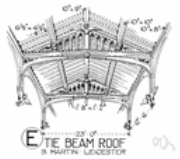
- Figuring out the exact meaning of a word or entity

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

Analysis with definitions(s)

*Bill Gates has developed an interest/***[readiness to give attention]** *in language technology and yesterday aquired a 10 % interest/***[a share (in a company, business, etc.)]** *in Torbjörn Lager 's sense disambiguation technology . Lager will retain a 90 % interest/***[a share (in a company, business, etc.)]** *in the new company , which will be based in Göteborg , Sweden . Last year 's drop in interest/***[money paid for the use of money]** *rates will probably be good for the company . Finally , although all this may sound like an arcane maneuver of little interest/***[quality of causing attention to be given]** *outside Wall Street , it would set off an economical earthquake .*

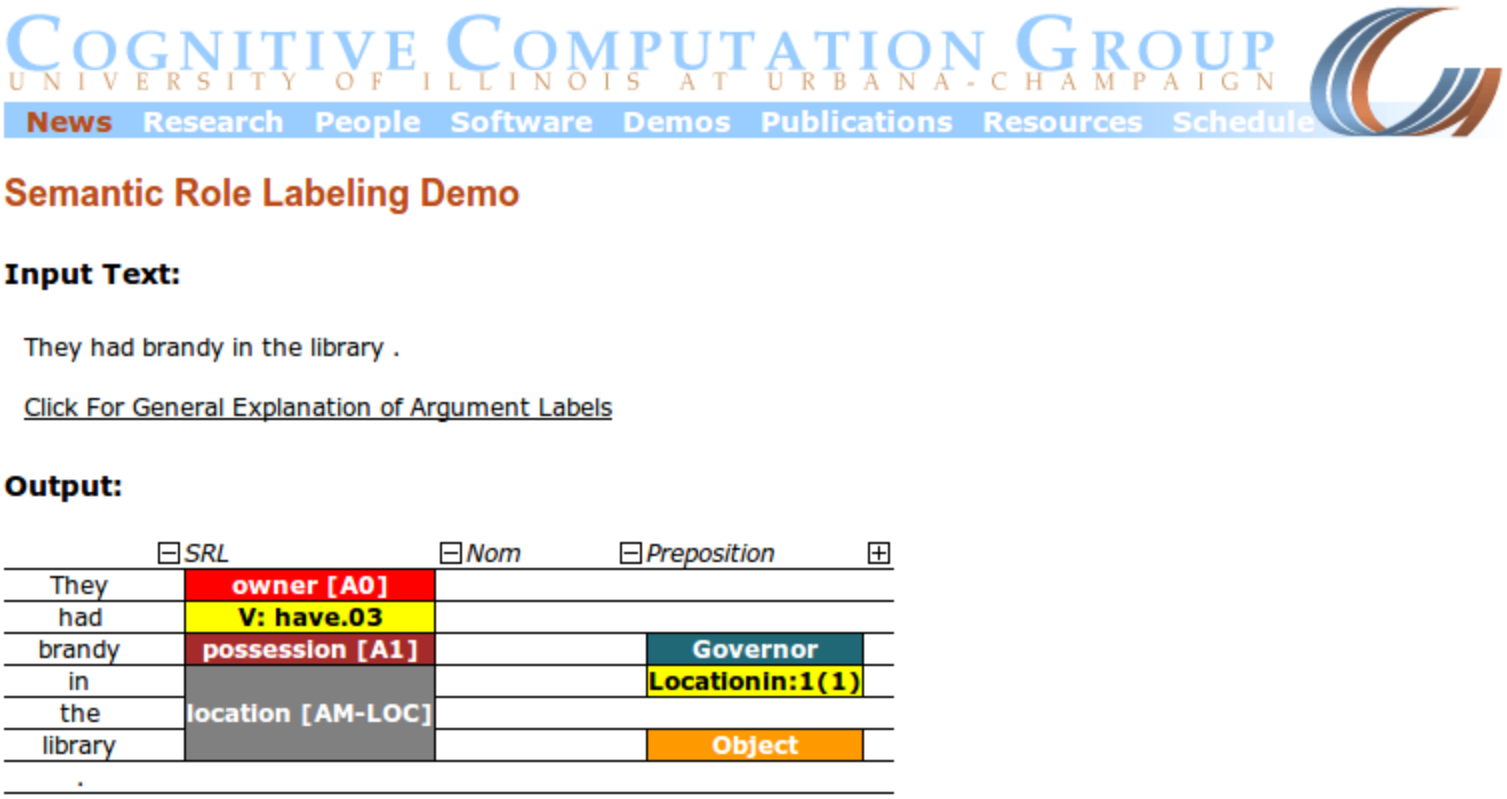
These are the six senses of the noun *interest* according to the LDOCE:

Sense	Definition
1	readiness to give attention
2	quality of causing attention to be given
3	activity, subject, etc., which one gives time and attention to
4	advantage, advancement, or favour
5	a share (in a company, business, etc.)
6	money paid for the use of money

http://www.ling.gu.se/~lager/Home/pwe_ui.html

Semantic role labeling

- Extracting subject-predicate-object triples from a sentence



http://cogcomp.cs.illinois.edu/page/demo_view/srl

Outline

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

Phonetics and phonology

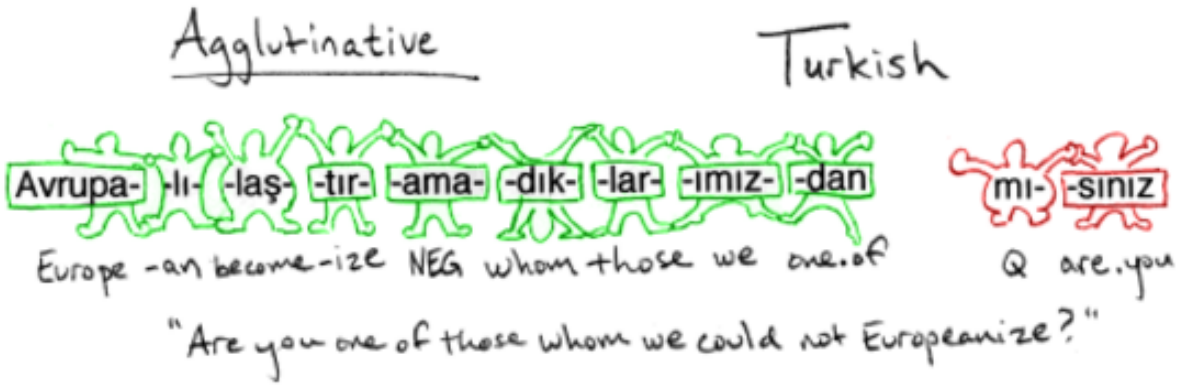
- The study of linguistic sounds and their relations to words

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



(http://allthingslinguistic.com/post/50939757945/morphological-typology-illustrations-from)

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): animal & dog
  - Meronymy (part of): finger & hand
  - Homonymy: fall (verb & season)
  - Antonymy: big & small

Pragmatics

- Social use of language
- 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!”

or

“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.

Berlin (/bərˈlɪn/, German: [bɛɐ̯ˈliːn] (listen)) is the capital of Germany, and one of the 16 states of Germany. With a population of 3.5 million people,[4] Berlin is Germany's largest city. It is the second most populous city proper and the seventh most populous urban area in the European Union.[5] Located in northeastern Germany on the banks of River Spree, it is the center of the Berlin-Brandenburg Metropolitan Region, which has about 6 million residents from over 180 nations.[6][7][8][9] Due to its location in the European Plain, Berlin is influenced by a temperate seasonal climate. Around one third of the city's area is composed of forests, parks, gardens, rivers and lakes.[10]

(http://en.wikipedia.org/wiki/Berlin)

Outline

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

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?

(http://www.realtytrac.com/landing/2009-year-end-foreclosure-report.html)

Semantics

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

(http://www.businessnewsdaily.com/2023-celebrity-hiring.html)

(http://en.wikipedia.org/wiki/Star#/media/File:Starsinthesky.jpg)

Discourse analysis

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

Outline

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

NLP Course

- Home page:
  - https://hpi.de/plattner/teaching/summer-term-2017/natural-language-processing.html
- Lecture
  - Monday 11:00-12:30
  - HS3
  - 3 credit points

Grading

- 60% Project
- 40% Final exam (written)

- You have to pass both of them

Program

(Program is subject to change)

Session	Date	Topic
1	April 24, 2017	Introduction to Natural Language Processing
2	May 1, 2017	(no lecture - Maifeiertag)
3	May 8, 2017	Morphology
4	May 15, 2017	Words and Language Model
5	May 22, 2017	Part-of-Speech Tagging and Syntactic Parsing
6	May 29, 2017	Lexical Semantics
7	June 5, 2017	(no lecture - Pfingstmontag)
8	June 12, 2017	Discourse analysis
9	June 19, 2017	Project: mid-term presentation (15 minutes for each team)
10	June 26, 2017	Sentiment Analysis
11	July 3, 2017	Named-entity Recognition and Information Extraction
12	July 10, 2017	Question Answering and Summarization
13	July 17, 2017	Project: final presentation (15 minutes for each team)
14	July 24, 2017	Final exam

Project

- Development of a NLP application
  - Information Retrieval
  - Information Extraction
  - Text Summarization
  - Question Answering
  - Sentiment Analysis
  - Machine Translation
  - Etc..

Project

- The application should include following components:
  - Syntactic parsing (e.g., POS tagging)
  - Semantics (e.g., NER)
  - Discourse analysis

Project

- Any NLP or ML libraries
  - Stanford Core NLP
  - NLTK
  - spaCy
  - Apache OpenNLP
  - GATE
  - SAP HANA (contact me)
  - R
  - Weka
  - TensorFlow
  - etc.

Project

- Any language
  - English, German, etc.

    (Check available NLP tools)
- Any text collections
  - Social media, Web pages, publications, Wikipedia, etc.
  - Benchmarks or newly created datasets
- Any domain

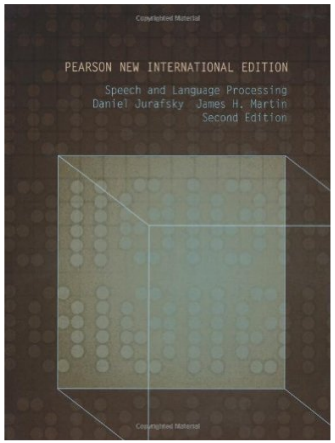
Project

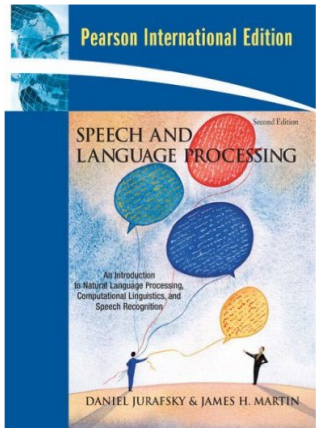
- Teams (2-3 students)
- Send me an email with your proposal (until May 5th, 2017)

- Updates (presentations) on the progress of the project
  - Mid-term and final presentation
  - Also considered for grading (commitment)

Course book

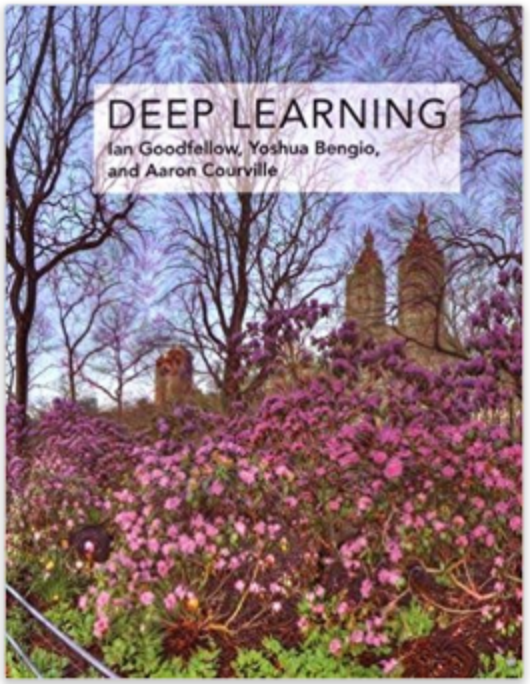
- Speech and Language Processing
  - Daniel Jurafsky and James H. Martin
  - New (3rd) edition: https://web.stanford.edu/~jurafsky/slp3/





Deep learning book

- Deep Learning (Adaptive Computation and Machine Learning)
  - Ian Goodfellow, Yoshua Bengio, Aaron Courville
  - The Mit Press, 2017
  - http://www.deeplearningbook.org
  - NLP in section 12.4



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

NLP Course

- Contact
  - Mariana.Neves@hpi.uni-potsdam.de
  - Room V-0.01 (Villa) - appointment under request

- We have a student position for NLP at the EPIC chair!