

Text Retrieval Based on Medical Subwords

Martin Honeck¹, Udo Hahn², Rüdiger Klar¹, Stefan Schulz¹

¹ Department of Medical Informatics
University Hospital Freiburg, Germany

² Natural Language Processing Division,
Freiburg University, Germany

Problem:

Poor performance of medical text retrieval in morphologically rich languages*

*most languages other than English

Linguistic Phenomena hamper Medical Text Retrieval

■ Word formation

(inflection, derivation, composition):

ulcus, ulcer^a, diagnosⁱs, diagnos^es, diagnostic, hep^ar, hep^atic, para|sympath|ectomy, proct|o|sigmoid|o|scop|ie, Rechts|herz|insuffizienz

■ Synonymy, spelling variants

{oesophagus, esophagus}, {leuko, leuco}, {Magenulcus, Magenulkus}, {cutis, skin}, {hemorrhage, bleeding}, {ascorbic, Vitamin C}, {ancylostoma, hookworm}

■ Multiple meanings:

Cold {low temperature, common cold}, Bruch {fracture, hernia}, APA {antiperoxidase antibodies, american psychology association}

Example

■ Frequency of German Word forms in *Google* Searches

Spelling Variants Synonyms			Inflections			Derivations		
Kolonkarzinom	2070	1780	Kolonkarzinom	2070	1770	Karzinom	17000	16900
Colonkarzinom	248	135	Kolonkarzinoms	471	253	karzinomatös	43	16
Colonicarcinom	111	73	Kolonkarzinome	275	139	karzinomatösen	86	40
Colon-Ca	203	169	Kolonkarzinomen	265	166	karzinomatöse	74	46
Kolon-Ca	66	46				karzinomatösem	7	5
Dickdarmkrebs	4000	3610				kazinomatöses	6	0
Dickdarmkarzinom	288	175				karzinomatöser	39	26
Dickdarmcarcinom	13	10						

Number of Hits

Number of exclusive hits (no other form matches)

Hypothesis:

Improving Text Retrieval Performance using
Linguistic Techniques

Subword as Index Terms for Text Retrieval

- Subwords are atomic linguistic sense units :
 - Morphemes: *nephr, anti, thyr, scler, hepat, cardi*
 - Morpheme aggregates: *diaphys, ascorb, anabol, diagnost*
 - Words: *amyloid, bone, fever, liver*
 - (noun groups: *vitamin c,...*)
- Criterion: well-defined, non-decomposable medical concepts
- Grouping of synonymous subwords:
 - kkyxkj** = {*nephr, kidney, nier, ren*},
 - qxxkjkq** = {*hepar, hepat, liver*},

Resources

- Subword lexicons:
Organize and classify subwords, prefixes and suffixes in several languages
- Subword thesaurus: Groups synonymous lexicon entries, links „similar“ groups
- Morphosyntactic parser: extracts subwords from text

Cf. Schulz et. al. MEDINFO 2001
Yearbook of Medical Informatics '02

Examples of Subword Extraction

■ Examples:

■ **proct** o **sigm** oid o **scop** y

■ **Schilddrüs** en **karzin** om

■ **cole cist ectom** ía

■ **acro cefal** o **sindattil** ía

■ **Sport verletz** ung en

■ **hør** sel s **hemm** ed e

■ **orchid** o **pex** ie

■ **Magen schleimhaut entzünd** ung

Lexical
subwords
(used for
indexing)

Functional
morphemes
(not used for
indexing)

Experiment:

Does Subword-based medical text retrieval
behave better than conventional methods ?
(formative evaluation - work in progress)

Retrieval Experiments: Sources

- German version of the ‘Merck Manual’ (medical textbook composed of 5,500 articles)
- 25 randomly chosen expert queries from medical students (German)
- 27 randomly chosen layman queries from the medical search engine “*Dr. Antonius*”
- Gold Standard:
Three medical students did manual relevance assessment (52 * 5,500 binary relevance judgements)

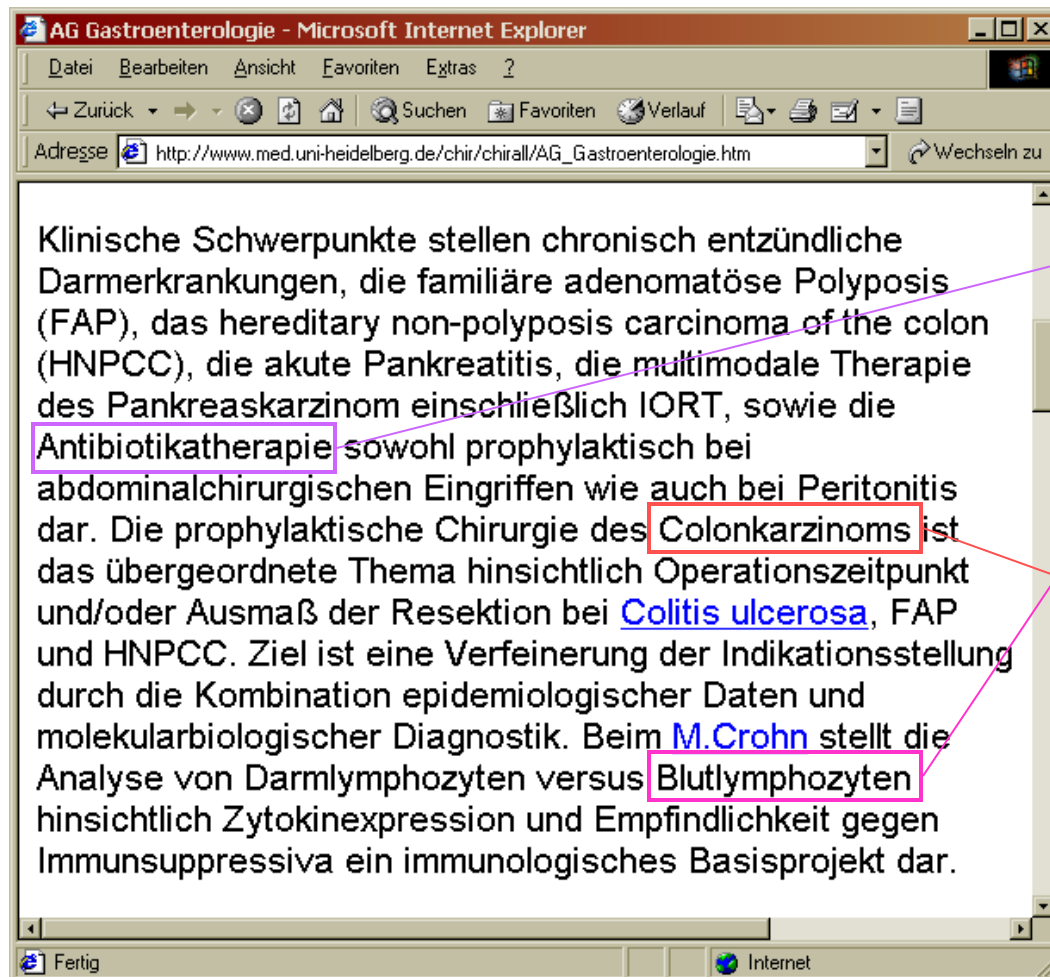
Retrieval Experiments:

- Salton's Vector Space Retrieval Engine (produces ranked output)
- Proximity boost (proximity of query terms in documents matters for document ranking)
- Tests:
 - Test 1 (plain): Token Search. Baseline
 - Test 2 (segm): Morphological Segmentation
 - Test 3 (norm): Morphological Segmentation and Synonym Expansion.

For all tests:

- Orthographic normalization preprocessing (e.g. ca → ka ,ci → zi, ä → ae, ...)

Token-based Indexing



abdominalchirurgischen
adenomatöse
akute
analyse
antibiotikatherapie
ausmaß
basisprojekt
blutlymphozyten
carcinoma
chirurgie
chronisch
colitis
colon
colonkarzinoms
darkerkrankungen
darmlymphozyten
daten
diagnostik
eingriffen
einschließlich
empfindlichkeit
entzündliche
epidemiologischen

Subword Indexing

AG Gastroenterologie - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Zurück Suchen Favoriten Verlauf Wechseln zu

Adresse http://www.med.uni-heidelberg.de/chir/chirall/AG_Gastroenterologie.htm

Klinische Schwerpunkte stellen chronisch entzündliche Darmerkrankungen, die familiäre adenomatöse Polyposis (FAP), das hereditary non-polyposis carcinoma of the colon (HNPCC), die akute Pankreatitis, die multimodale Therapie des Pankreaskarzinom einschließlich IORT, sowie die Antibiotikatherapie sowohl prophylaktisch bei abdominalchirurgischen Eingriffen wie auch bei Peritonitis dar. Die prophylaktische Chirurgie des Colonkarzinoms ist das übergeordnete Thema hinsichtlich Operationszeitpunkt und/oder Ausmaß der Resektion bei Colitis ulcerosa, FAP und HNPCC. Ziel ist eine Verfeinerung der Indikationsstellung durch die Kombination epidemiologischer Daten und molekularbiologischer Diagnostik. Beim M.Crohn stellt die Analyse von Darmlymphozyten versus Blutlymphozyten hinsichtlich Zytokinexpression und Empfindlichkeit gegen Immunsuppressiva ein immunologisches Basisprojekt dar.

Fertig Internet

abdomin
adenom
akut
analys
antibiot
ausmass
basis
biolog
blut
chirurg
chroni
darm
daten
diagnost
eingriff
empfindlich
entzuend
epidemiolog
express
famili
fap
fein
heredit
hinsichtlich
hnpcc
immun
indik
iort
itis
karzin
klin
kolitis
kolon

kombin
krank
krohn
lymph
modal
molekul
multi
non
operation
ordn
osis
pankreas
pankreat
periton
polyp
projekt
prophylakt
punkt
resekt
schwerpunkt
stell
suppress
thema
therap
ueber
ulzer
versus
zeit
ziel
zyt
zytokin

Subword - Indexing with Semantic Normalization

AG Gastroenterologie - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Zurück Suchen Favoriten Verlauf Wechseln zu

Adresse http://www.med.uni-heidelberg.de/chir/chirall/AG_Gastroenterologie.htm

Klinische Schwerpunkte stellen chronisch **entzündliche** Darmerkrankungen, die familiäre adenomatöse Polyposis (FAP), das hereditary non-polyposis carcinoma of the colon (HNPCC), die akute **Pankreatitis**, die multimodale Therapie des Pankreaskarzinom einschließlich IORT, sowie die Antibiotikatherapie sowohl prophylaktisch bei abdominalchirurgischen Eingriffen wie auch bei **Peritonitis** dar. Die prophylaktische Chirurgie des Colonkarzinoms ist das übergeordnete Thema hinsichtlich Operationszeitpunkt und/oder Ausmaß der Resektion bei [Colitis ulcerosa](#), FAP und HNPCC. Ziel ist eine Verfeinerung der Indikationsstellung durch die Kombination epidemiologischer Daten und molekularbiologischer Diagnostik. Beim [M.Crohn](#) stellt die Analyse von Darmlymphozyten versus Blutlymphozyten hinsichtlich Zytokinexpression und Empfindlichkeit gegen Immunsuppressiva ein immunologisches Basisprojekt dar.

Fertig Internet

qxxqky
yxyqwx
yyxqkx
zzkqy
yyzq
kkqkky
qkqzzk
yzxqk
qxqkxz
kqxqqk
kzzkqz
yzqyyz
yzkkzy
xqkzqq
yqqqkq
xxzxqk
zxkqqq
qyyyzx
kzxqkk
kqkzzq
kqqzkz
yzqkqz
zzqqzz
yyyyyq
kkqyzq
qqkqzz
kqkyzy
yqqkkk
kxyzqk
zxqkyz
kkzqxy
qqkqkz

zzyqkk
yzxqkz
zzkqz
zzkqzqz
yyyzyk
ykzyqk
xzqqqz
qkqkqz
zxqkyy
xkqqqy
yyyzxk
zxqkkq
qkqzqz

{entzuend; inflamm;
itis}

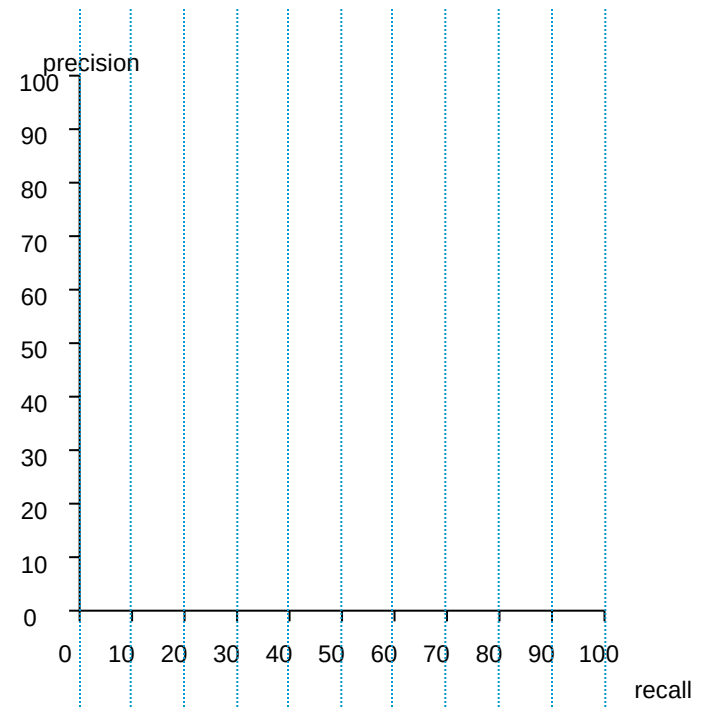
{pankreas; pankreat;
bauchspeicheldrues}

{periton; bauchfell}

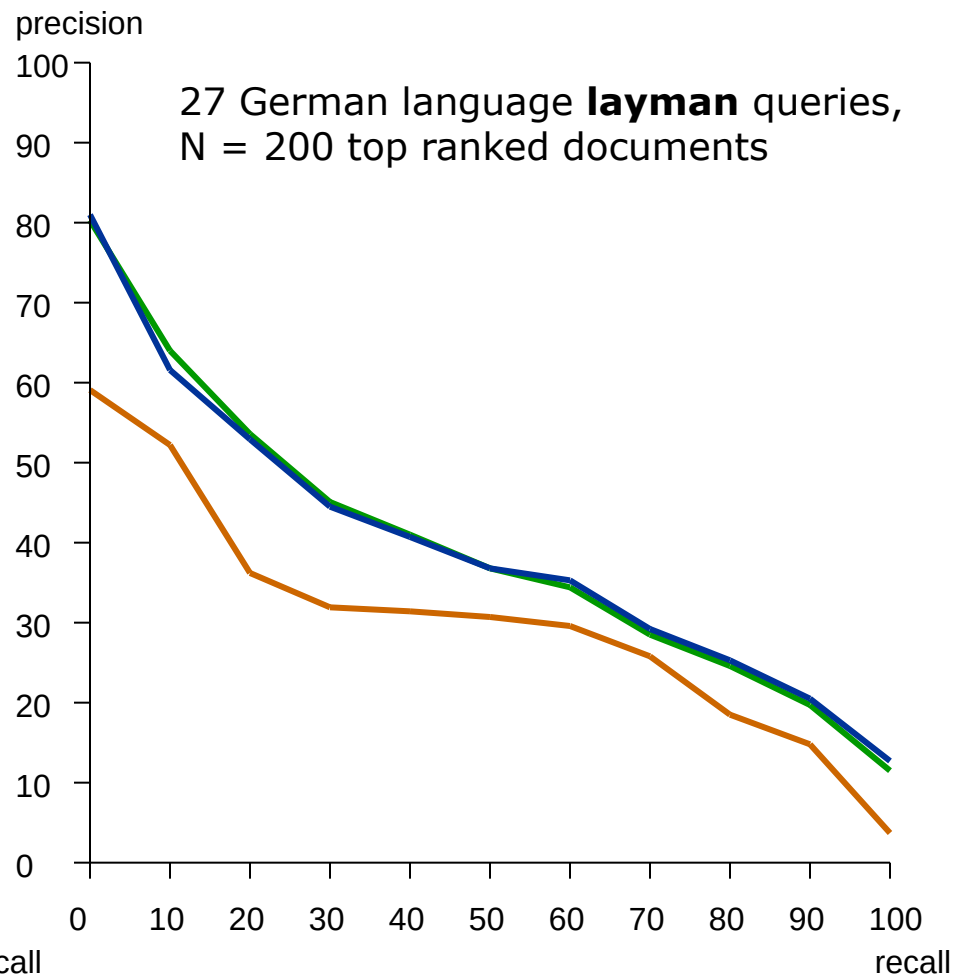
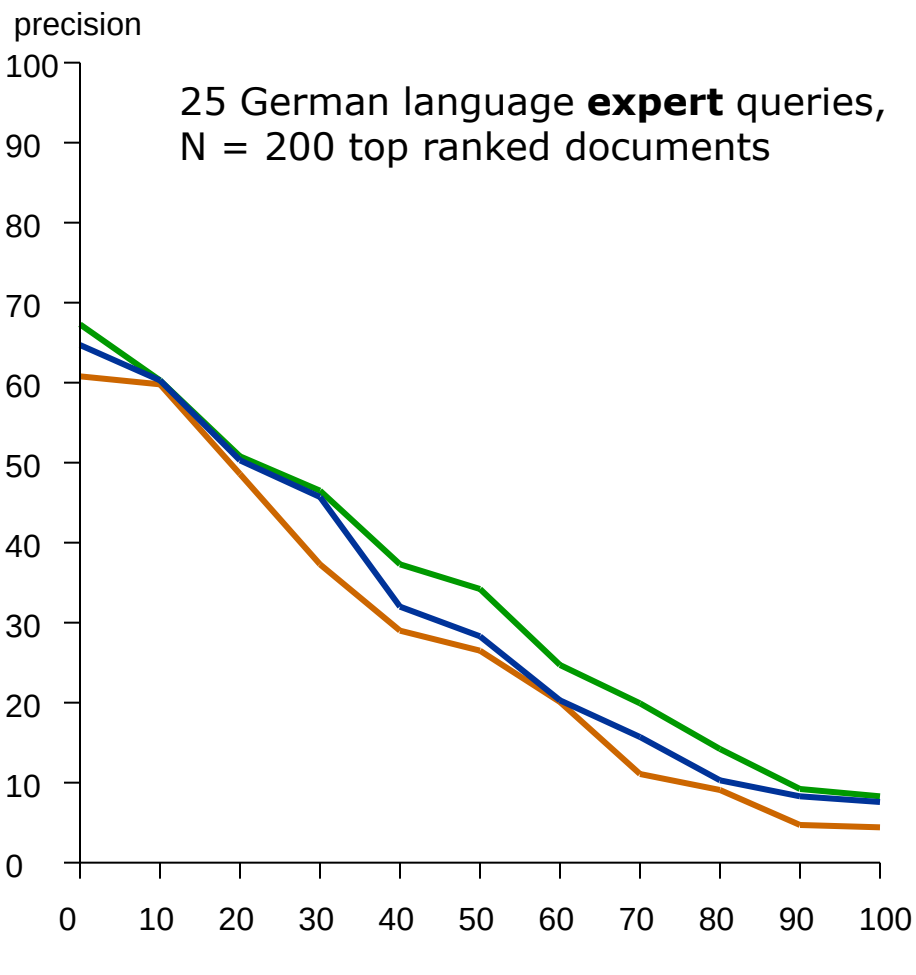
kyzykq
qkkkyq
xyzqkq
qkqkqy

Presentation of Results

- Precision / Recall Diagrams
- For each query:
interpolation of precision
value at fixed recall levels
(0%, 10%, ..., 100%)
- Arithmetic mean of
precision values at each
recall level



Retrieval Experiments: Results



- Test 1: Token Search ("plain"). Baseline
- Test 2: Morphological Segmentation ("segm")
- Test 3: Morphological Segmentation and Synonym Expansion. ("norm").

Significance Judgements

Recall (%)	Precision (%)								
	<i>plain</i>	<i>segm</i>	<i>norm</i>	<i>plain</i>	<i>segm</i>	<i>norm</i>	<i>plain</i>	<i>segm</i>	<i>norm</i>
	expert queries n=25			layman queries n=27			all queries n=52		
0	60.8	67.3	64.7	59.1	80.3	81.0	60.0	74.0	73.2
10	59.8	60.3	60.3	52.2	64.0	61.6	55.8	62.3	61.0
20	48.6	50.8	50.3	36.2	53.6	52.9	42.1	52.3	51.7
30	37.3	46.5	45.7	31.9	45.1	44.5	34.5	45.8	45.1
40	29.0	37.3	32.0	31.4	41.0	40.7	30.3	39.2	36.5
50	26.5	34.2	28.3	30.7	36.8	36.8	28.7	35.6	32.7
60	20.1	24.7	20.3	29.6	34.4	35.3	25.0	29.7	28.1
70	11.1	19.9	15.7	25.8	28.5	29.2	18.7	24.4	22.7
80	9.1	14.2	10.3	18.5	24.6	25.3	14.0	19.6	18.1
90	4.7	9.2	8.3	14.8	19.7	20.5	9.9	14.7	14.6
100	4.4	8.3	7.6	3.7	11.5	12.7	4.0	10.0	10.2
11pt avrg.	24.1	33.9	31.2	29.5	40.0	40.0	26.9	37.0	35.8

$\alpha < 0.05$ (Wilcoxon test)

Discussion:

Do the results justify the effort ?

Discussion

- Work in progress
- Coverage of Subword dictionary (core vocabulary of clinical medicine (excl. proper names, acronyms) for German, English, Portuguese, ~ 17,000 entries). Target: 30,000 entries
- Linking subwords by synonymy relations adds noise to the system: more cautious use of synonymy relation
- Noise due to the erroneous extraction of medical subwords from non-medical terms and proper names: inclusion in dictionary

Outlook

- Data-driven improvement of lexicons, thesaurus word grammar, algorithms, disambiguation heuristics
- Automated acquisition of abbreviations and acronyms (WWW)
- Semi-Automated acquisition of proper names
- Linkage to (MeSH): concept hierarchies, synonyms at the level of noun groups
- Evaluation of monolingual retrieval for Portuguese
- Evaluation of cross-lingual retrieval (German - English, English - Portuguese)

Evaluation of Text Retrieval Systems

■ Target variables:

$$precision = \frac{n_{found+relevantDocuments}}{n_{found_documents}}$$

$$recall = \frac{n_{found+relevant_documents}}{n_{relevant_documents}}$$

■ Precision/Recall-Diagrams with ranked output

Example: 25 documents, 8 relevant

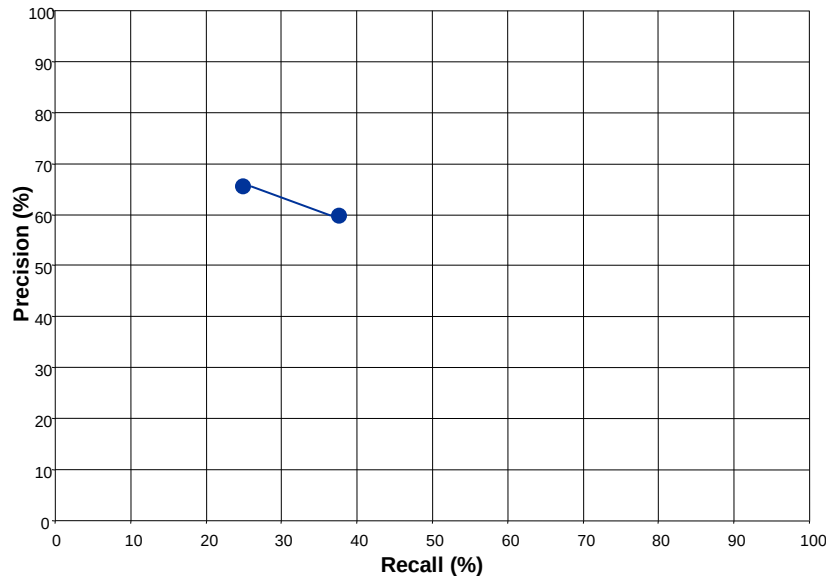
precision = 67%
recall = 25%

Query X



document 05
document 16
document 21
document 22
document 02
document 25
document 20
document 10
document 07
document 18
document 04
document 12
document 11
document 24
document 15
document 09
document 17
document 08
document 19
document 13
document 03
document 14
document 23
document 01
document 06

Evaluation of Text Retrieval Systems



■ Precision/Recall-Diagrams with ranked output

Example: 25 documents, 8 relevant

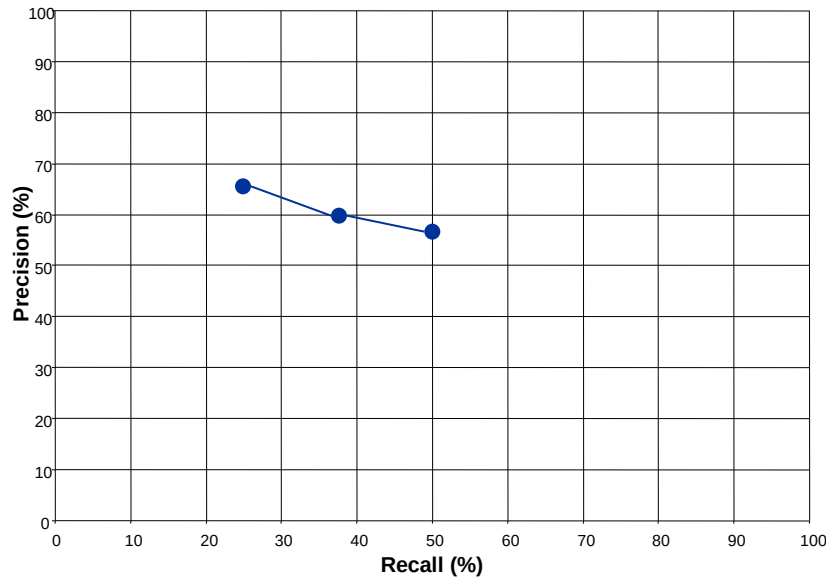
Query X



precision = 60%
recall = 38%

document 05
document 16
document 21
document 22
document 02
document 25
document 20
document 10
document 07
document 18
document 04
document 12
document 11
document 24
document 15
document 09
document 17
document 08
document 19
document 13
document 03
document 14
document 23
document 01
document 06

Evaluation of Text Retrieval Systems



■ Precision/Recall-Diagrams with ranked output

Example: 25 documents, 8 relevant

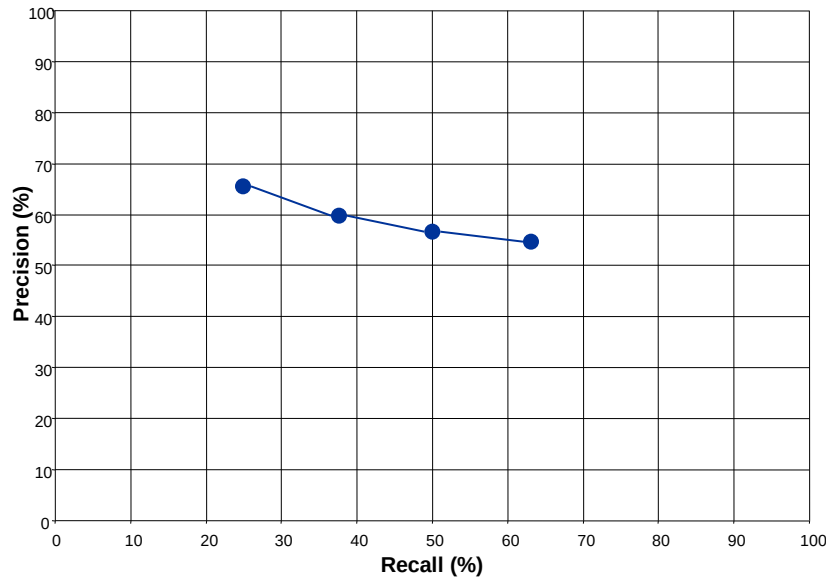
Query X



precision = 57%
recall = 50%

document 05
document 16
document 21
document 22
document 02
document 25
document 20
document 10
document 07
document 18
document 04
document 12
document 11
document 24
document 15
document 09
document 17
document 08
document 19
document 13
document 03
document 14
document 23
document 01
document 06

Evaluation of Text Retrieval Systems



■ Precision/Recall-Diagrams with ranked output

Example: 25 documents, 8 relevant

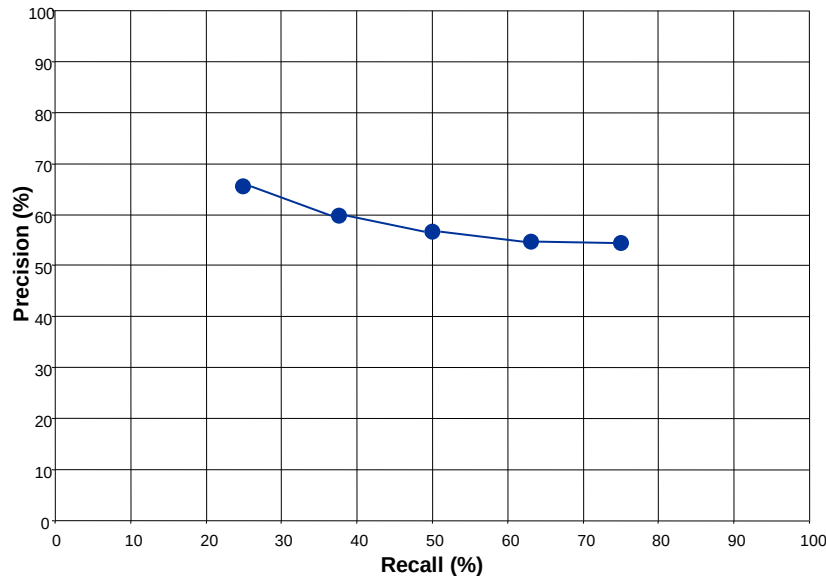
Query X



precision = 55%
recall = 63%

document 05
document 16
document 21
document 22
document 02
document 25
document 20
document 10
document 07
document 18
document 04
document 12
document 11
document 24
document 15
document 09
document 17
document 08
document 19
document 13
document 03
document 14
document 23
document 01
document 06

Evaluation of Text Retrieval Systems



■ Precision/Recall-Diagrams with ranked output

Example: 25 documents, 8 relevant

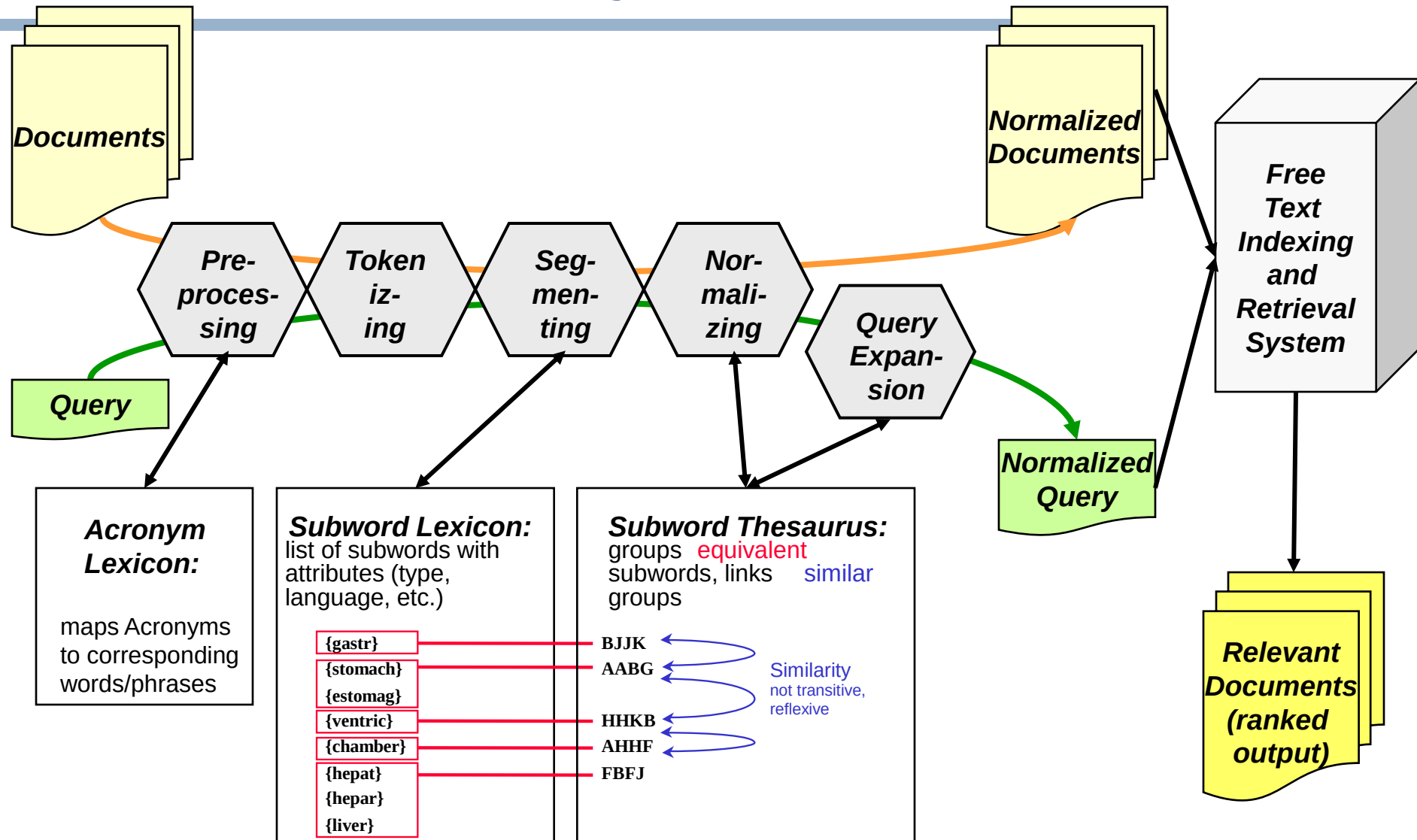
Query X



precision = 54%
recall = 75%

document 05
document 16
document 21
document 22
document 02
document 25
document 20
document 10
document 07
document 18
document 04
document 12
document 11
document 24
document 15
document 09
document 17
document 08
document 19
document 13
document 03
document 14
document 23
document 01
document 06

Extended System Architecture



Tool: Subword Editor & Workbench

editing tool for
subword lexicon
and thesaurus

Segmentation

Please enter text here:

abdominal pain

Segmented text:

ABDOMIN -al PAIN

Active Language
☐ German ☒ English

Db Update Update + Segment Segment

morpheme thesaurus construction tool [C:\Dokumente\mck neu 008\thetemplate1909200...

File Morphemes Thesaurus ?

Lexem A	Type	Weight	Lang
abdomen	1	3	1

a
ab
abdomen
abdomin
abduzens
abduz
abel
aberr
abfall
abfluss
abfuhr
abgeleitet
abhaeng

☐ Find
☐ Match

☒ left truncation

Sort / Filter A

EqClass A 4

abdomen
abdomin
unterleib

abdomen
5205
belly
5206
abdomin

SIM

deSIM

EQ

deEQ A

AutoEQ

?

Lexem B	Type	Weight	Lang
belly	1	3	2

belly
abdomin
abduzens
abdukt
able
aberr
arthropod
artificial
like
artikul
arytaen
aryten
drug

☐ Find
☐ Match

☒ left truncation

Sort / Filter B

EqClass B 5205

belly

4
abdomen
abdomin
unterleib
340
bauch

Morphemes: 9462 EqClasses: 9064 SIM-Rel: 6604 Compactness f(r/c) .73 Net Density f(r/c) 1.04

testbed for
segmentation

The Subword Approach (II)

- Language-specific algorithms for **extraction** of subwords from (medical) texts
- **Multilingual** subword **repositories**
- Criteria for subword delimitation and classification
 - Semantic (compositionality)
Hyper | cholesterol | emia
 - Lexical (enabling synonym matching)
schleimhaut = mucosa (~~schleim | haut~~)
 - Data-driven (avoiding ambiguities and false segmentation), e.g.
relation~~ship~~, Schwanger~~schaft~~ (~~relation | ship~~, ~~Schwanger | schaft~~)

Disfunção tireoideana perinatal

As doenças da tireóide acometem 10% das mulheres, mas a maioria das pacientes responde bem ao tratamento.

Durante a gestação, mudanças metabólicas podem ocultar a patologia, com risco de dano fetal devido à conduta inapropriada. Exames de TSH, tiroxina livre e triiodotironina livre são essenciais.

Geralmente, a presença de valores elevados de TSH sugere o diagnóstico de hipotireoidismo primário, enquanto níveis suprimidos de TSH sugerem hipertireoidismo. Este último costuma manifestar-se através de bócio, oftalmopatia, fraqueza muscular, taquicardia ou perda de peso.

DIS FUNCAO TIREOID e ana PERI NATAL

as DOENCA s da TIREOID e ACOMET em 10% das MULHER es MAS a MAIOR ia das PACIENT es RESPOND e BEM ao TRATAMENT o.

DURANTE a GESTAC ao MUDANCA s METABOL ic as PRESENC a da PATOLOG ia COM RISC o de DAN o a CONDUT a in APROPRIAD a. os EXAME s de “TSH”, e TRI IODO TIRONIN a LIVR e sao ESSENCI ais GERAL mente a PRESENC a de VALOR es ELEVAD os de “TSH” SUGER e o DIAGNOST ic o de HIPO TIREOID ism o PRIMAR io ENQUANTO NIVEIS SUPRIM id os de “TSH” SUGER em HIPER TIREOID ism o. este ULTIM o COSTUM a MANIFEST ar se ATRAVES de BOCIO, OFTALM o PATIA FRAQU eza MUSCUL ar TAQUI CARD ia ou PERD a de PESO.

iiiiil iiifunct iiithyr iiibirth

iiipatho iiithyr iiiaffect 10% iiifemin iiibut iiigh iiipatient iiirepond iiigood iitreatment.

iiiduring iipregnan iichange iimetabol iipossibl ii with iirisk iidamage iifetus iidue iibehav iisuita iihormon, iiithyroxin iifree iithree iijod iiithyronin iiigeneral iipresent iivalue iiigh iiithyr iistimul iil iiidiagnos iilow iiithyr iifirst iiduring iilevel iisuppress iiithyr iistimul iihormon iisuggest iiigh iiithyrii. iilast iicustom iimanifest iiby iigoiter, iiieye iipatho iiweak iimuscule iispeed iiheart iilose iiweigh.

Perinatal Thyroid Dysfunction

Thyroid gland diseases affect 10% of women, but most patients respond well to treatment.

, metabolic changes can hide the presence of the risk of fetal damage due to inappropriate handling. SH, free T4 and T3 are indispensable.

Generally, high TSH values suggest the diagnosis of primary hypothyroidism while a suppressed TSH level suggests hyperthyroidism. Typical manifestations of the latter are goiter, opthalmopathy, muscular weakness, tachycardia, or weight loss

PERI NATAL THYROID DYS FUNCTION

THYROID GLAND DISEAS es AFFECT 10% of WOMEN BUT MOST PATIENT s RESPOND WELL to TREATMENT

cy METABOL ic CHANGE s CAN HIDE the DISORDER WITH the RISK of FETAL DAMAGE DUE to HANDL ing. MEASURE ment of “TSH”, FREE “T4” and “T3” are INDISPENSABLE GENERAL ly HIGH “TSH” VALUE s SUGGEST the DIAGNOS is of PRIMAR y HYPO THYROID ism WHILE a SUPPRESS ed “TSH” LEVEL SUGGEST s HYPER THYROID ism. TYP ic al MANIFEST ation s of the LATTER are GOITER, OPHTALM o PATHY, MUSCUL ar WEAK ness TACHY CARD y or WEIGHT LOSS.

iiibirth iiithyr iiiiil iiifunct

iiithyr iiigland iiipatho iiiaffect 10% iiifemin iiibut iiigh iiipatient iiirepond iiigood iitreatment

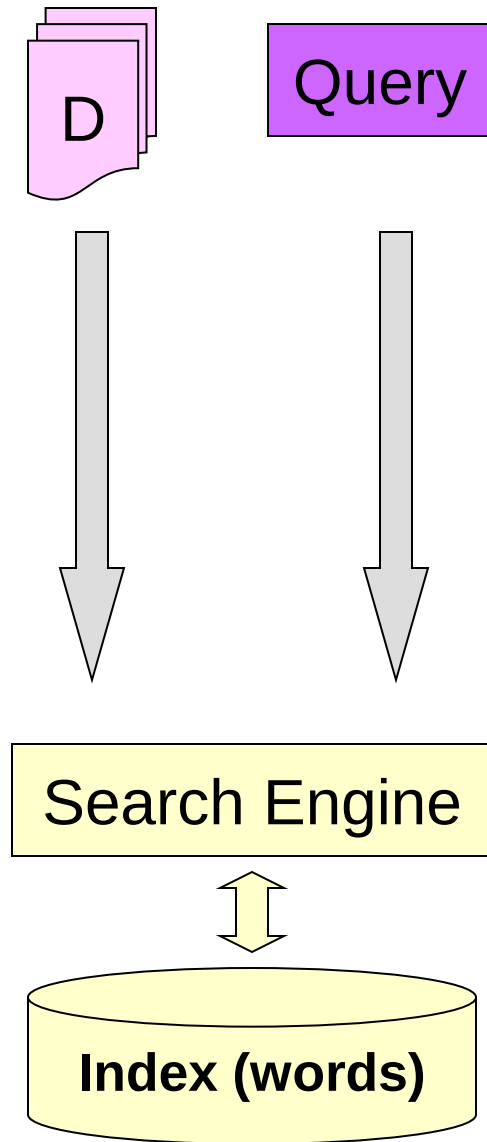
ol iichange iican iihide iipresent iipatho iiwith iiisuitabl iimanag. iiimeasure iiithyr iistimul iithree iijod iiithyronin iiessential mul iihormon iivalue iisuggest iiidiagnos iifirst iilow iiithyr iiduring iisuppress iiithyr iistimul iihormon iilevel iisuggest iiigh iiithyr. iityp iimanifest iilast iiigoiteriii, iiieye iipathoiii, iimuscule iiweak iispeed iiheart iiweigh iilose..

Original text (D)

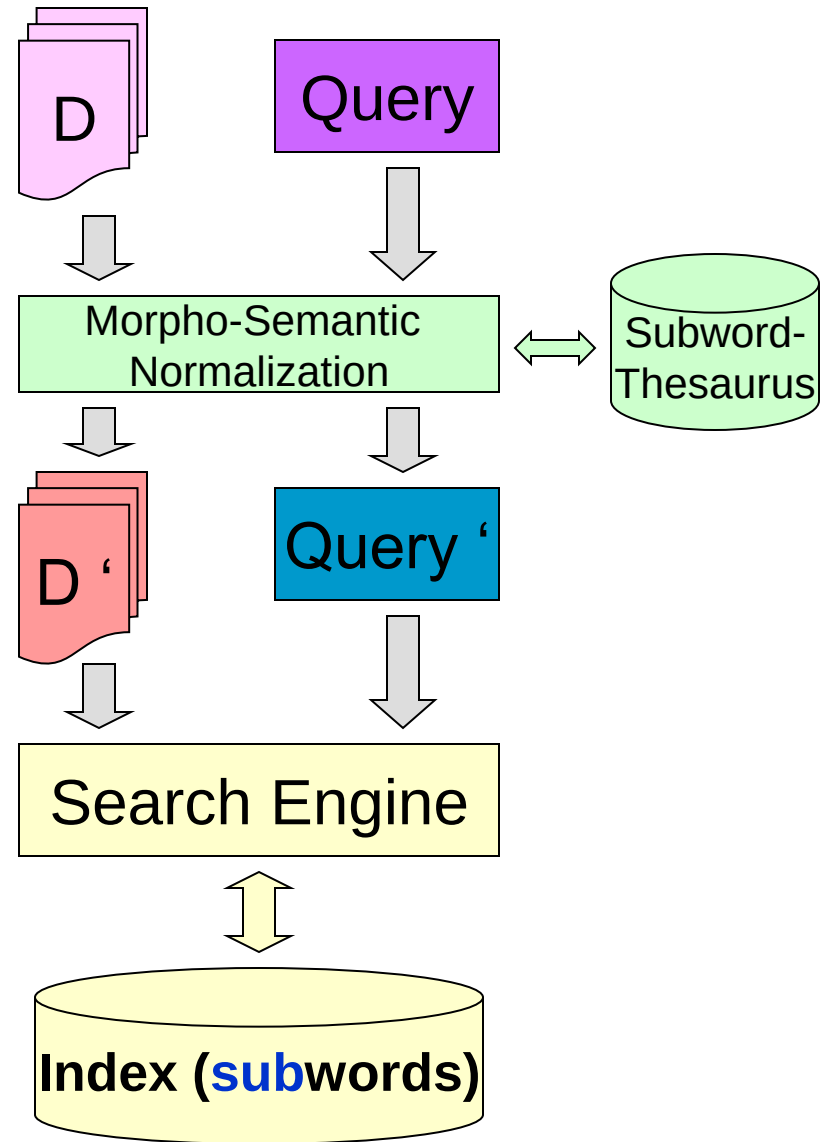
Segmented text

Segmented text mapped to thesaurus Ids (D')

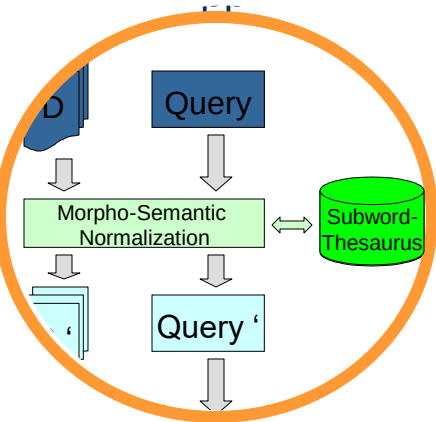
Conventional approach



Subword approach



Lexical Resources



Subword Lexicon:

list of subwords with attributes (type, language, etc.)

{gastr}

{stomach}

{magen}

{ventric}

{chamber}

{hepat}

{hepar}

{liver}

{kidney}

{ren}

{nier}

Equivalence
transitive
and reflexive

Subword Thesaurus:

groups **equivalent** subwords,
links similar groups

ID#

ykzyqk

jkzyqj

zyzzjj

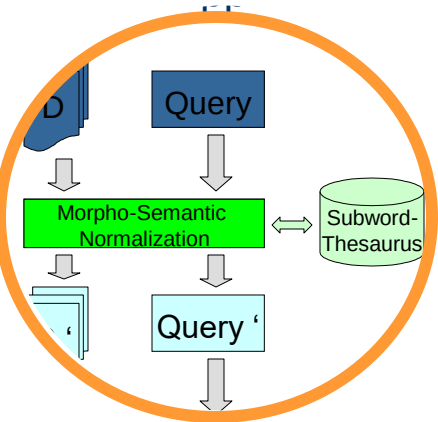
xjkkkq

qxkjkq

kkyxkj

Similarity
not transitive,
reflexive

Algorithmic Resources



- Morphosyntactic parser based on a word model described as a finite-state automaton
- Heuristic rules for disambiguation of parses

