PLIN021 Sémantická analýza v praxi

OP VK Mezi bohemistikou a informatikou
www.projekt-inova.cz

Zuzana Nevěřilová
xpopelk@fi.muni.cz

Centrum zpracování přirozeného jazyka, B203
Fakulta informatiky, Masarykova univerzita

6. května 2013

Významové potenciály

Zkoumání kontextu

Kontextové vektory

Odbočka k teorii množin

Matematický model významu

Významové potenciály (Hanks)

- slovníky pro lidi jsou pro počítače nevyhovující
- slovníky kombinované a hierarchické (WordNet) mají lepší výsledky v počítačovém zpracování (Nirenburg 2007)
- význam není jedno políčko v „kontrolním seznamu“
- významy lze nejlépe interpretovat pomocí pravděpodobnosti, s jakou se užití blíží prototypu (Fillmore)
- významy jsou spojeny se vzory (patterns, Hunston a Francis, 2000), konstrukcemi (Goldberg) či „frazémy“ (Melčuk), ale i se slovy – korpusová lingvistika dokáže tyto vzory (preference užití) zjistit a studovat
- významy vně kontextu neexistují, existují jen významové potenciály, které se kontextem aktivují [Hanks, 2000]

Významové potenciály (Hanks)

- slovníky pro lidi jsou pro počítače nevyhovující
- slovníky kombinované a hierarchické (WordNet) mají lepší výsledky v počítačovém zpracování (Nirenburg 2007)
- význam není jedno políčko v „kontrolním seznamu“
- významy lze nejlépe interpretovat pomocí pravděpodobnosti, s jakou se užití blíží prototypu (Fillmore)
- významy jsou spojeny se vzory (patterns, Hunston a Francis, 2000), konstrukcemi (Goldberg) či „frazémy“ (Melčuk), ale i se slovy – korpusová lingvistika dokáže tyto vzory (preference užití) zjistit a studovat
- významy vně kontextu neexistují, existují jen významové potenciály, které se kontextem aktivují [Hanks, 2000]

Významové potenciály (Hanks)

- slovníky pro lidi jsou pro počítače nevyhovující
- slovníky kombinované a hierarchické (WordNet) mají lepší výsledky v počítačovém zpracování (Nirenburg 2007)
- význam není jedno políčko v „kontrolním seznamu“
- významy lze nejlépe interpretovat pomocí pravděpodobnosti, s jakou se užití blíží prototypu (Fillmore)
- významy jsou spojeny se vzory (patterns, Hunston a Francis, 2000), konstrukcemi (Goldberg) či „frazémy“ (Melčuk), ale i se slovy – korpusová lingvistika dokáže tyto vzory (preference užití) zjistit a studovat
- významy vně kontextu neexistují, existují jen významové potenciály, které se kontextem aktivují [Hanks, 2000]

Významové potenciály (Hanks)

- slovníky pro lidi jsou pro počítače nevyhovující
- slovníky kombinované a hierarchické (WordNet) mají lepší výsledky v počítačovém zpracování (Nirenburg 2007)
- význam není jedno políčko v „kontrolním seznamu“
- významy lze nejlépe interpretovat pomocí pravděpodobnosti, s jakou se užití blíží prototypu (Fillmore)
- významy jsou spojeny se vzory (patterns, Hunston a Francis, 2000), konstrukcemi (Goldberg) či „frazémy“ (Melčuk), ale i se slovy – korpusová lingvistika dokáže tyto vzory (preference užití) zjistit a studovat
- významy vně kontextu neexistují, existují jen významové potenciály, které se kontextem aktivují [Hanks, 2000]

Významové potenciály (Hanks)

- slovníky pro lidi jsou pro počítače nevyhovující
- slovníky kombinované a hierarchické (WordNet) mají lepší výsledky v počítačovém zpracování (Nirenburg 2007)
- význam není jedno políčko v „kontrolním seznamu“
- významy lze nejlépe interpretovat pomocí pravděpodobnosti, s jakou se užití blíží prototypu (Fillmore)
- významy jsou spojeny se vzory (patterns, Hunston a Francis, 2000), konstrukcemi (Goldberg) či „frazémy“ (Melčuk), ale i se slovy – korpusová lingvistika dokáže tyto vzory (preference užití) zjistit a studovat
- významy vně kontextu neexistují, existují jen významové potenciály, které se kontextem aktivují [Hanks, 2000]

Významové potenciály (Hanks)

- slovníky pro lidi jsou pro počítače nevyhovující
- slovníky kombinované a hierarchické (WordNet) mají lepší výsledky v počítačovém zpracování (Nirenburg 2007)
- význam není jedno políčko v „kontrolním seznamu“
- významy lze nejlépe interpretovat pomocí pravděpodobnosti, s jakou se užití blíží prototypu (Fillmore)
- významy jsou spojeny se vzory (patterns, Hunston a Francis, 2000), konstrukcemi (Goldberg) či „frazémy“ (Melčuk), ale i se slovy – korpusová lingvistika dokáže tyto vzory (preference užití) zjistit a studovat
- významy vně kontextu neexistují, existují jen významové potenciály, které se kontextem aktivují [Hanks, 2000]

Významové potenciály (Hanks)

Corpus Patterns Analysis

PATTERN: [[Human]] translate ([[Document]]) (from [[Language 1]]) (into [[Language 2]])

IMPLICATURE: [[Human]] expresses the meaning of [[Document]] in [[Language 1]] in the words and phraseology of [[Language 2]]

Významové potenciály (Hanks): normy (norm) [Hanks, 2010]

- norma = užití slova podle určitého syntagmatického vzoru
- slova se užívají jednak v souladu s normou (očekáváním), jednak mluvčí normu porušují
- vzory se skládají z užitého slova a z lexikálních množin, se kterými se slovo užívá
- lexikální množiny mohou být obrovské, např. [[Human]]
- čím menší lexikální množina, tím silnější je její vliv na význam vzoru

Významové potenciály (Hanks): normy (norm) [Hanks, 2010]

- norma = užití slova podle určitého syntagmatického vzoru
- slova se užívají jednak v souladu s normou (očekáváním), jednak mluvčí normu porušují
- vzory se skládají z užitého slova a z lexikálních množin, se kterými se slovo užívá
- lexikální množiny mohou být obrovské, např. [[Human]]
- čím menší lexikální množina, tím silnější je její vliv na význam vzoru

Významové potenciály (Hanks): normy (norm) [Hanks, 2010]

- norma = užití slova podle určitého syntagmatického vzoru
- slova se užívají jednak v souladu s normou (očekáváním), jednak mluvčí normu porušují
- vzory se skládají z užitého slova a z lexikálních množin, se kterými se slovo užívá
- lexikální množiny mohou být obrovské, např. [[Human]]
- čím menší lexikální množina, tím silnější je její vliv na význam vzoru

Významové potenciály (Hanks): normy (norm) [Hanks, 2010]

- norma = užití slova podle určitého syntagmatického vzoru
- slova se užívají jednak v souladu s normou (očekáváním), jednak mluvčí normu porušují
- vzory se skládají z užitého slova a z lexikálních množin, se kterými se slovo užívá
- lexikální množiny mohou být obrovské, např. [[Human]]
- čím menší lexikální množina, tím silnější je její vliv na význam vzoru

Významové potenciály (Hanks): normy (norm) [Hanks, 2010]

- norma = užití slova podle určitého syntagmatického vzoru
- slova se užívají jednak v souladu s normou (očekáváním), jednak mluvčí normu porušují
- vzory se skládají z užitého slova a z lexikálních množin, se kterými se slovo užívá
- lexikální množiny mohou být obrovské, např. [[Human]]
- čím menší lexikální množina, tím silnější je její vliv na význam vzoru

Významové potenciály (Hanks): porušení normy (exploitation)

- kreativní užití jazyka
- porušení normy nepřesahuje 10 % případů v korpusu
- pokud ano, je to nejspíš dosud neobjevená norma
- i porušení normy má jistá pravidla „dvojitá šroubovice“ systémů pravidel: pravidla, jak slova používat normálně, a pravidla, jak normu porušit
- často studovaným vzorem je valence

Významové potenciály (Hanks): porušení normy (exploitation)

- kreativní užití jazyka
- porušení normy nepřesahuje 10 % případů v korpusu
- pokud ano, je to nejspíš dosud neobjevená norma
- i porušení normy má jistá pravidla „dvojitá šroubovice“ systémů pravidel: pravidla, jak slova používat normálně, a pravidla, jak normu porušit
- často studovaným vzorem je valence

Významové potenciály (Hanks): porušení normy (exploitation)

- kreativní užití jazyka
- porušení normy nepřesahuje 10 % případů v korpusu
- pokud ano, je to nejspíš dosud neobjevená norma
- i porušení normy má jistá pravidla „dvojitá šroubovice“ systémů pravidel: pravidla, jak slova používat normálně, a pravidla, jak normu porušit
- často studovaným vzorem je valence

Významové potenciály (Hanks): porušení normy (exploitation)

- kreativní užití jazyka
- porušení normy nepřesahuje 10 % případů v korpusu
- pokud ano, je to nejspíš dosud neobjevená norma
- i porušení normy má jistá pravidla „dvojitá šroubovice“ systémů pravidel: pravidla, jak slova používat normálně, a pravidla, jak normu porušit
- často studovaným vzorem je valence

Významové potenciály (Hanks): porušení normy (exploitation)

- kreativní užití jazyka
- porušení normy nepřesahuje 10 % případů v korpusu
- pokud ano, je to nejspíš dosud neobjevená norma
- i porušení normy má jistá pravidla „dvojitá šroubovice“ systémů pravidel: pravidla, jak slova používat normálně, a pravidla, jak normu porušit
- často studovaným vzorem je valence

Významové potenciály (Hanks): valence

...možná bych prodloužila poslední verš...	nesedí	mi tam nejen vizuálně, že je příliš krátký
Whitesun a stále nedokončeným C-3PO. Cliegg	sedí	na létajícím křesle, protože mu chybí jedna
atraktivní módní přehlídce. Tvůrkyně ošacení	seděla	na premiéře v první řadě v atraktivní róbě
zablokovali oba směry magistrály. Někteří	sedí	na schodech k opeře. Mávají rudými vlajkami
centrem byl zdemolován vůz TV Nova. Jinak zde	sedí	ještě menší skupina demonstrantů. Jde většinou
hráli v sále hospody a hodně lidí zůstalo	sedět	venku, protože tentokrát počasí přálo.
jsem byl naprosto unešena.Jsou perfekní.	Sedí	jako ulité.Jen délka ,ale tu jsem za 15
psala, dorazil a věcičky jsou nádherný Vše	sedí	tak jak má a já jsem moooc spokojená Moc
dnes mi s podprsenkami dorazil... Skvěle	sedí	. Jsem moc spokojená! Budu se těšit zase
moc děkuji, těhotenské rifle jsou super -	sedí	úplně perfektně!!! A taky děkuju za dáreček

Významové potenciály (Hanks): valence

...možná bych prodloužila poslední verš...	nesedí	mi tam nejen vizuálně, že je příliš krátký
Whitesun a stále nedokončeným C-3PO. Cliegg	sedí	na létajícím křesle, protože mu chybí jedna
atraktivní módní přehlídce. Tvůrkyně ošacení	seděla	na premiéře v první řadě v atraktivní róbě
zablokovali oba směry magistrály. Někteří	sedí	na schodech k opeře. Mávají rudými vlajkami
centrem byl zdemolován vůz TV Nova. Jinak zde	sedí	ještě menší skupina demonstrantů. Jde většinou
hráli v sále hospody a hodně lidí zůstalo	sedět	venku, protože tentokrát počasí přálo.
jsem byl naprosto unešena.Jsou perfekní.	Sedí	jako ulité.Jen délka ,ale tu jsem za 15
psala, dorazil a věcičky jsou nádherný Vše	sedí	tak jak má a já jsem moooc spokojená Moc
dnes mi s podprsenkami dorazil... Skvěle	sedí	. Jsem moc spokojená! Budu se těšit zase
moc děkuji, těhotenské rifle jsou super -	sedí	úplně perfektně!!! A taky děkuju za dáreček

[[Human | Group of Humans]] sedět [[Location]]
[[Garment]] sedět [[Human, dative]] [[Manner]]
[[Art]] sedět [[Human, dative]] [[Location]] [[Manner]]

Významové potenciály (Hanks): elipsa

...možná bych prodloužila poslední verš...	nesedí	mi tam nejen vizuálně, že je příliš krátký
Whitesun a stále nedokončeným C-3PO. Cliegg	sedí	na létajícím křesle, protože mu chybí jedna
atraktivní módní přehlídce. Tvůrkyně ošacení	seděla	na premiéře v první řadě v atraktivní róbě
zablokovali oba směry magistrály. Někteří	sedí	na schodech k opeře. Mávají rudými vlajkami
centrem byl zdemolován vůz TV Nova. Jinak zde	sedí	ještě menší skupina demonstrantů. Jde většinou
hráli v sále hospody a hodně lidí zůstalo	sedět	venku, protože tentokrát počasí přálo.
jsem byl naprosto unešena.Jsou perfekní.	Sedí	jako ulité.Jen délka ,ale tu jsem za 15
psala, dorazil a věcičky jsou nádherný Vše	sedí	tak jak má a já jsem moooc spokojená Moc
dnes mi s podprsenkami dorazil... Skvěle	sedí	. Jsem moc spokojená! Budu se těšit zase
moc děkuji, těhotenské rifle jsou super -	sedí	úplně perfektně!!! A taky děkuju za dáreček

[[Human | Group of Humans]] sedět [[Location]]
[[Garment]] sedět [[Human, dative]] [[Manner]]
[[Art]] sedět [[Human, dative]] [[Location]] [[Manner]]

Významové potenciály (Hanks): elipsa

...možná bych prodloužila poslední verš...	nesedí	mi tam nejen vizuálně, že je příliš krátký
Whitesun a stále nedokončeným C-3PO. Cliegg	sedí	na létajícím křesle, protože mu chybí jedna
atraktivní módní přehlídce. Tvůrkyně ošacení	seděla	na premiéře v první řadě v atraktivní róbě
zablokovali oba směry magistrály. Někteří	sedí	na schodech k opeře. Mávají rudými vlajkami
centrem byl zdemolován vůz TV Nova. Jinak zde	sedí	ještě menší skupina demonstrantů. Jde většinou
hráli v sále hospody a hodně lidí zůstalo	sedět	venku, protože tentokrát počasí přálo.
jsem byl naprosto unešena.Jsou perfekní.	Sedí	jako ulité.Jen délka ,ale tu jsem za 15
psala, dorazil a věcičky jsou nádherný Vše	sedí	tak jak má a já jsem moooc spokojená Moc
dnes mi s podprsenkami dorazil... Skvěle	sedí	. Jsem moc spokojená! Budu se těšit zase
moc děkuji, těhotenské rifle jsou super -	sedí	úplně perfektně!!! A taky děkuju za dáreček

sedět [[Location]]

sedět [[Manner]]

sedět [[Human, dative]] [[Location]]

Významové potenciály (Hanks): elipsa

Elipsy mají také svá pravidla. Co může být vypuštěno?

- podmět
- předmět
- příslovečné určení

Za jakých okolností se může vyskytnout elipsa? Vypuštěno může být jen to, co je zřejmé.

Elipsa jako porušení normy? Je to otázka frekvence.

Kontextové vektory [Schütze, 1998]

Významy jsou spojeny vztahy.

Kontextové vektory [Schütze, 1998]

Významy jsou spojeny vztahy.

Zdá se, že některé významy jsou „víc spojeny“ než jiné. Např. „pták“ je víc spojený s „peří“ než se „strom“.

Kontextové vektory [Schütze, 1998]

Významy jsou spojeny vztahy.

Zdá se, že některé významy jsou „víc spojeny“ než jiné. Např. „pták“ je víc spojený s „peří“ než se „strom“.

Problémem WSD je inventář významů, jeho kvalita, granularita a aktuálnost. Inventářům se můžeme vyhnout, pokud potřebujeme „pouze“ zjistit, která slova jsou použita ve stejném významu, aniž bychom věděli, jaký význam to je.

Kontextové vektory [Schütze, 1998]

Významy jsou spojeny vztahy.

Zdá se, že některé významy jsou „víc spojeny“ než jiné. Např. „pták“ je víc spojený s „peří“ než se „strom“.

Problémem WSD je inventář významů, jeho kvalita, granularita a aktuálnost. Inventářům se můžeme vyhnout, pokud potřebujeme „pouze“ zjistit, která slova jsou použita ve stejném významu, aniž bychom věděli, jaký význam to je.

Algoritmus rozlišení kontextových skupin (context group discrimination)

Výsledkem jsou výskyty víceznačného slova v různých shlucích. Každé slovo, kontext i shluk jsou reprezentovány vektorem v mnoharozměrném vektorovém prostoru.

Vektor jako reprezentant výskytu slova v doméně

Mějme n domén $d_i \in \mathcal{D} | i = 1, \ldots, n$ (např. zoologie, vaření, atmosféra, vojenské letectví).

Každé slovo w je reprezentováno vektorem $v = (x_1, x_2, \ldots, x_n)$.

Vyskytuje-li se slovo w v textech z domény d_i, pak x_i přiřadíme četnost w v doméně d_i.

Četnost můžeme vyjádřit více způsoby (které už známe z WSD):

- počet výskytů w
- počet dokumentů, ve kterých se w vyskytuje
- 0 pokud se w v d_i nevyskytuje, jinak 1
- ...

Vektor jako reprezentant výskytu slova v doméně

Mějme 4 domény $d_i \in \mathcal{D} | i = 1, \ldots, 4$ (zoologie, vaření, atmosféra, vojenské letectví).

Každé slovo w je reprezentováno vektorem $v = (x_1, x_2, x_3, x_4)$.

Získáme potom vektory:

$v_1(\text{buňka}) = (10, 0, 0, 5)$ $\quad |v_1| = \sqrt{100 + 25} = 11,18$

$v_2(\text{tkáň}) = (9, 0, 0, 0)$ $\quad |v_2| = \sqrt{81} = 9$

$v_3(\text{let}) = (4, 0, 1, 10)$ $\quad |v_3| = \sqrt{16 + 1 + 100} = 10,81$

$v_4(\text{množství}) = (4, 5, 4, 5)$ $\quad |v_4| = \sqrt{16 + 25 + 16 + 25} = 9,06$

$v_5(\text{pára}) = (0, 6, 5, 1)$ $\quad |v_5| = \sqrt{36 + 25 + 1} = 7,87$

Vektor jako reprezentant výskytu slova v doméně

$v_1(\text{buňka}) = (10, 0, 0, 5)$ $\quad |v_1| = 11,18$
$v_2(\text{tkáň}) = (9, 0, 0, 0)$ $\quad |v_2| = 9$
$v_3(\text{let}) = (4, 0, 1, 10)$ $\quad |v_3| = 10,81$
$v_4(\text{množství}) = (4, 5, 4, 5)$ $\quad |v_4| = 9,06$
$v_5(\text{pára}) = (0, 6, 5, 1)$ $\quad |v_5| = 7,87$

$$\arccos(v_1, v_2) = \arccos \frac{v_1 \cdot v_2}{|v_1| \cdot |v_2|} = \arccos \frac{10 \cdot 9 + 0 \cdot 0 + 0 \cdot 0 + 5 \cdot 0}{11,18 \cdot 9} =$$
$$\arccos \frac{90}{100,62} = \arccos 0,89 = 27°$$

α	v_1	v_2	v_3	v_4	v_5
v_1	0	$27°$	$42,2°$	$50°$	$86,6°$

Vektor jako reprezentant výskytu slova v doméně

$v_1(\text{buňka}) = (10, 0, 0, 5)$ $\quad |v_1| = 11,18$
$v_2(\text{tkáň}) = (9, 0, 0, 0)$ $\quad |v_2| = 9$
$v_3(\text{let}) = (4, 0, 1, 10)$ $\quad |v_3| = 10,81$
$v_4(\text{množství}) = (4, 5, 4, 5)$ $\quad |v_4| = 9,06$
$v_5(\text{pára}) = (0, 6, 5, 1)$ $\quad |v_5| = 7,87$

$$\arccos(v_1, v_2) = \arccos \frac{v_1 \cdot v_2}{|v_1| \cdot |v_2|} = \arccos \frac{10 \cdot 9 + 0 \cdot 0 + 0 \cdot 0 + 5 \cdot 0}{11,18 \cdot 9} =$$
$$\arccos \frac{90}{100,62} = \arccos 0,89 = 27°$$

α	v_1	v_2	v_3	v_4	v_5
v_1	0	27°	42,2°	50°	86,6°
v_2	27°	0	68°	63,9°	90°

Vektor jako reprezentant výskytu slova v doméně

$v_1(\text{buňka}) = (10, 0, 0, 5)$ $\quad |v_1| = 11,18$
$v_2(\text{tkáň}) = (9, 0, 0, 0)$ $\quad |v_2| = 9$
$v_3(\text{let}) = (4, 0, 1, 10)$ $\quad |v_3| = 10,81$
$v_4(\text{množství}) = (4, 5, 4, 5)$ $\quad |v_4| = 9,06$
$v_5(\text{pára}) = (0, 6, 5, 1)$ $\quad |v_5| = 7,87$

$$\arccos(v_1, v_2) = \arccos \frac{v_1 \cdot v_2}{|v_1| \cdot |v_2|} = \arccos \frac{10 \cdot 9 + 0 \cdot 0 + 0 \cdot 0 + 5 \cdot 0}{11,18 \cdot 9} =$$
$$\arccos \frac{90}{100,62} = \arccos 0,89 = 27°$$

α	v_1	v_2	v_3	v_4	v_5
v_1	0	27°	42,2°	50°	86,6°
v_2	27°	0	68°	63,9°	90°
v_3	42,2°	68°	0	44,4°	80°

Vektor jako reprezentant výskytu slova v doméně

$v_1(\text{buňka}) = (10, 0, 0, 5)$ $\quad |v_1| = 11,18$
$v_2(\text{tkáň}) = (9, 0, 0, 0)$ $\quad |v_2| = 9$
$v_3(\text{let}) = (4, 0, 1, 10)$ $\quad |v_3| = 10,81$
$v_4(\text{množství}) = (4, 5, 4, 5)$ $\quad |v_4| = 9,06$
$v_5(\text{pára}) = (0, 6, 5, 1)$ $\quad |v_5| = 7,87$

$$\arccos(v_1, v_2) = \arccos \frac{v_1 \cdot v_2}{|v_1| \cdot |v_2|} = \arccos \frac{10 \cdot 9 + 0 \cdot 0 + 0 \cdot 0 + 5 \cdot 0}{11,18 \cdot 9} =$$
$$\arccos \frac{90}{100,62} = \arccos 0,89 = 27°$$

α	v_1	v_2	v_3	v_4	v_5
v_1	0	27°	42,2°	50°	86,6°
v_2	27°	0	68°	63,9°	90°
v_3	42,2°	68°	0	44,4°	80°
v_4	50°	63,9°	44,4°	0	40°

Vektor jako reprezentant výskytu slova v doméně

$v_1(\text{buňka}) = (10, 0, 0, 5)$ $\quad |v_1| = 11,18$
$v_2(\text{tkáň}) = (9, 0, 0, 0)$ $\quad |v_2| = 9$
$v_3(\text{let}) = (4, 0, 1, 10)$ $\quad |v_3| = 10,81$
$v_4(\text{množství}) = (4, 5, 4, 5)$ $\quad |v_4| = 9,06$
$v_5(\text{pára}) = (0, 6, 5, 1)$ $\quad |v_5| = 7,87$

$$\arccos(v_1, v_2) = \arccos \frac{v_1 \cdot v_2}{|v_1| \cdot |v_2|} = \arccos \frac{10 \cdot 9 + 0 \cdot 0 + 0 \cdot 0 + 5 \cdot 0}{11,18 \cdot 9} =$$
$$\arccos \frac{90}{100,62} = \arccos 0,89 = 27°$$

α	v_1	v_2	v_3	v_4	v_5
v_1	0	27°	42,2°	50°	86,6°
v_2	27°	0	68°	63,9°	90°
v_3	42,2°	68°	0	44,4°	80°
v_4	50°	63,9°	44,4°	0	40°
v_5	86,6°	90°	80°	40°	0

Kontextové vektory [Schütze, 1998]

Algoritmus:

1. vytvoř matici spoluvýskytů
2. spočítej kontextový vektor pro každý kontext
3. sdruž kontextové vektory do shluků

[Král, 2006]

Kontextové vektory [Schütze, 1998]

$v_1(\text{buňka}) = (10, 0, 0, 5)$
$v_2(\text{tkáň}) = (9, 0, 0, 0)$
$v_3(\text{let}) = (4, 0, 1, 10)$
$v_4(\text{množství}) = (4, 5, 4, 5)$
$v_5(\text{pára}) = (0, 6, 5, 1)$

w	zoologie	vaření	atmosféra	vojenské letectví
buňka	10	0	0	5
tkáň	9	0	0	0
let	4	0	1	10
množství	4	5	4	5
pára	0	6	5	1

Kontextové vektory: matice spoluvýskytů

Matice spoluvýskytů je tabulka, kde řádky odpovídají znakům a sloupce dimenzím. Čísla v buňkách odpovídají počtu spoluvýskytu znaku a dimenze v tomtéž kontextu.

Kontextové vektory: matice spoluvýskytů

Matice spoluvýskytů je tabulka, kde řádky odpovídají znakům a sloupce dimenzím. Čísla v buňkách odpovídají počtu spoluvýskytu znaku a dimenze v tomtéž kontextu.

Jaká slova vybrat jako znaky? Ideálně všechna, většina z nich nebude mít žádný vliv (v buňkách budou nuly) a budeme je moci vypustit.

Kontextové vektory: matice spoluvýskytů

Matice spoluvýskytů je tabulka, kde řádky odpovídají znakům a sloupce dimenzím. Čísla v buňkách odpovídají počtu spoluvýskytu znaku a dimenze v tomtéž kontextu.

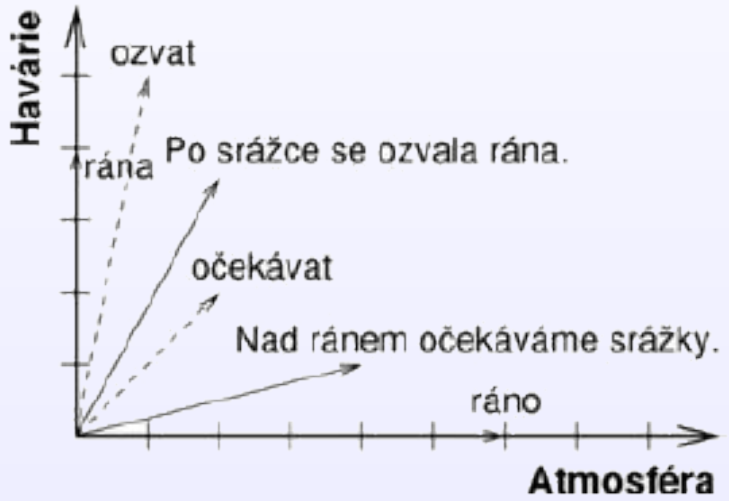
Jaká slova vybrat jako znaky? Ideálně všechna, většina z nich nebude mít žádný vliv (v buňkách budou nuly) a budeme je moci vypustit.

	Atmosféra *Atmosphere*	*Havárie* *Crash*
ráno/bang	0	4
ráno/morning	6	0
ozvat/resound	1	5
očekávat/expect	2	2

Kontextové vektory: výpočet

Kontextový vektor získáme jako průměrný vektor všech výskytů všech slov v daném kontextu.

	Atmosféra *Atmosphere*	*Havárie* *Crash*
rána/bang	0	4
ráno/morning	6	0
ozvat/resound	1	5
očekávat/expect	2	2



Po srážce se ozvala(1,5) rána $\frac{(0,4)+(6,0)}{2} = (3, 2)$.
$\frac{(1,5)+(3,2)}{2} = (2, 3\frac{1}{2})$

Nad ránem(6,0) očekáváme(2,2) srážky. $\frac{(6,0)+(2,2)}{2} = (4, 1)$

Kontextové vektory: shlukování (klastrování)

1. Vyber k centroidů (těžišť)
2. Každý kontextový vektor přiřaď k nejbližšímu centroidu
3. Centroid přepočítej podle přítomných vektorů
4. Opakuj kroky 2–3, dokud se shluky neustálí

Výsledek pro slovo *srážka*:

Cluster:	ozbrojený armed	daň tax	mzda wage	teplota temperature	oblačnost cloudiness	zahynout deaden	voják soldier	vlak train
1	0.07	0.01	0.01	0.01	0.00	0.02	0.04	0.01
2	0.01	0.16	0.20	0.01	0.00	0.01	0.01	0.00
3	0.02	0.01	0.01	0.12	0.08	0.01	0.01	0.00
4	0.03	0.01	0.01	0.01	0.00	0.08	0.02	0.04

Množina, n-tice, relace, zobrazení, funkce

Množina $A = \{x_1, \ldots, x_n\}$ soubor prvků. Množina je určena svými prvky. Množiny mohou být prvky jiných množin.

Součin $A \times B$ je množina (uspořádaných) dvojic.
$A \times B = \{(a, b) | a \in A, b \in B\}$

N-tice $(x_1, \ldots, x_n)$ je prvek součinu $A_1 \times \cdots \times A_n$, kde $x_i \in A_i$

Relace R je podmnožina součinu $A_1 \times \cdots \times A_n$

Funkce je relace $f \subset A \times B$, kde pro $x \in A$ existuje právě jedno $y \in B$ takové, že $(x, y) \in f$.

Zobrazení je obecnější než funkce. Funkce je zobrazení do množiny čísel.

Matematický model významu [Widdows, 2003]

přesné vymezení toho, co je kontext

prostory $\mathcal{W}$ (words), $\mathcal{L}$ (lexicon of meanings), $\mathcal{C}$ (contexts)

korespondence $(w, c) \rightarrow l$

kontextové skupiny: homonyma jsou v disjunktních kontextových skupinách, víceznačná slova jsou v překrývajících se k. skupinách

Matematický model významu: motivace

Soutěže jako SENSEVAL ukázaly, že úspěch či neúspěch WSD záleží na tom, jak těžké víceznačnosti jsou. Co to ale znamená?

Někdy mají potíže s rozeznáním významu i lidé, jak to pak mají zvládnout počítače?

Problém je jednak granularita, jednak kontext. Ve většině přístupů je totiž kontext definován vágně.

Matematický model významu: prostory

$\mathcal{W}$ (words)	slova, části složených slov, víceslovné výrazy

Matematický model významu: prostory

$\mathcal{W}$ (words)	slova, části složených slov, víceslovné výrazy
$\mathcal{L}$ (lexicon of meanings)	tradiční slovníky, ontologie, taxonomie, významy z trénovacích dat

Matematický model významu: prostory

$\mathcal{W}$ (words)	slova, části složených slov, víceslovné výrazy
$\mathcal{L}$ (lexicon of meanings)	tradiční slovníky, ontologie, taxonomie, významy z trénovacích dat
$\mathcal{C}$ (contexts)	věty, kolokace, domény

Matematický model významu: tradiční WSD

tradiční WSD: $(w, c) \in \mathcal{W} \times \mathcal{C}$

zobrazení: $\phi : (w, c) \to \mathcal{L}$

ověření oproti „zlatému standardu“ (tj. manuálním anotacím)

všechny významy slova: $S(w) = \{\phi(w, c) | \forall c \in \mathcal{C}\} \subset \mathcal{L}$

úkol WSD je extrapolace (zobecnění) ϕ
(známe hodnotu $\phi(w, c_1)$, odhadujeme $\phi(w, c_2)$)

Matematický model významu: synonymie

slova $w_1, w_2 \in \mathcal{W}$ jsou synonyma právě, když $\phi(w_1, c) = \phi(w_2, c)$

zobrazení z W do L není injektivní

úplná synonymie: $\phi(w_1, c) = \phi(w_2, c)$ pro všechna $c \in \mathcal{C}$

Matematický model významu: odposlouchávání

Odposlouchávání (eavesdropping) v neznámých datech: přiřazení významu nejen z kontextu $c \in \mathcal{C}$, ale z libovolné podmnožiny $\mathcal{C}$. Označme $\mathcal{C}_s$ kontexty, které jsou relevantní pro w. Pak přiřazení významu je zobrazení

$$\phi : (w, c, \mathcal{C}_s) \to \mathcal{L}$$

Jak zjistit $\mathcal{C}_s$? Pomocí podobností spočítaných na korpusu.

Matematický model významu: kontextové skupiny

Jak vlastně vypadá množina $\mathcal{C}$?

Matematický model významu: kontextové skupiny

Jak vlastně vypadá množina $\mathcal{C}$?

Podmnožina promluvy .....

Matematický model významu: kontextové skupiny

Jak vlastně vypadá množina $\mathcal{C}$?

Podmnožina promluvy . . . . . obsahující slovo w

Kolik kontextu potřebujeme pro určení významu w? Záleží případ od případu.

Matematický model významu: kontextové skupiny

Jak vlastně vypadá množina $\mathcal{C}$?

Podmnožina promluvy .....obsahující slovo w

Kolik kontextu potřebujeme pro určení významu w? Záleží případ od případu.

Tradiční přístup ke kontextu je $c = (w_1, \ldots, w_n)$, tj. zobrazení $\mathcal{W} \times \cdots \times \mathcal{W} = \mathcal{W}^n \to \mathcal{C}$

Matematický model významu: kontextové skupiny

Jak vlastně vypadá množina $\mathcal{C}$?

Podmnožina promluvy .....obsahující slovo w

Kolik kontextu potřebujeme pro určení významu w? Záleží případ od případu.

Tradiční přístup ke kontextu je $c = (w_1, \ldots, w_n)$, tj. zobrazení $\mathcal{W} \times \cdots \times \mathcal{W} = \mathcal{W}^n \to \mathcal{C}$

$\mathcal{W}$ je však širší, obsahuje „meta“ informace, obecně nepopsa(tel)né slovy, např.:

„v lékařském kontextu *operace* vždy znamená chirurgický zákrok, na rozdíl od vojenské nebo matematické operace“

Matematický model významu: kontextové skupiny

Vztah mezi významy a kontexty je monotónní, tj. jsou-li dva významy velmi různé, jsou velmi různé i kontexty, ve kterých se slovo objevuje.

Nabízí se tedy popsat vztah mezi významy a kontexty bez ohledu na to, jak *nějaký konkrétní kontext vypadá*.

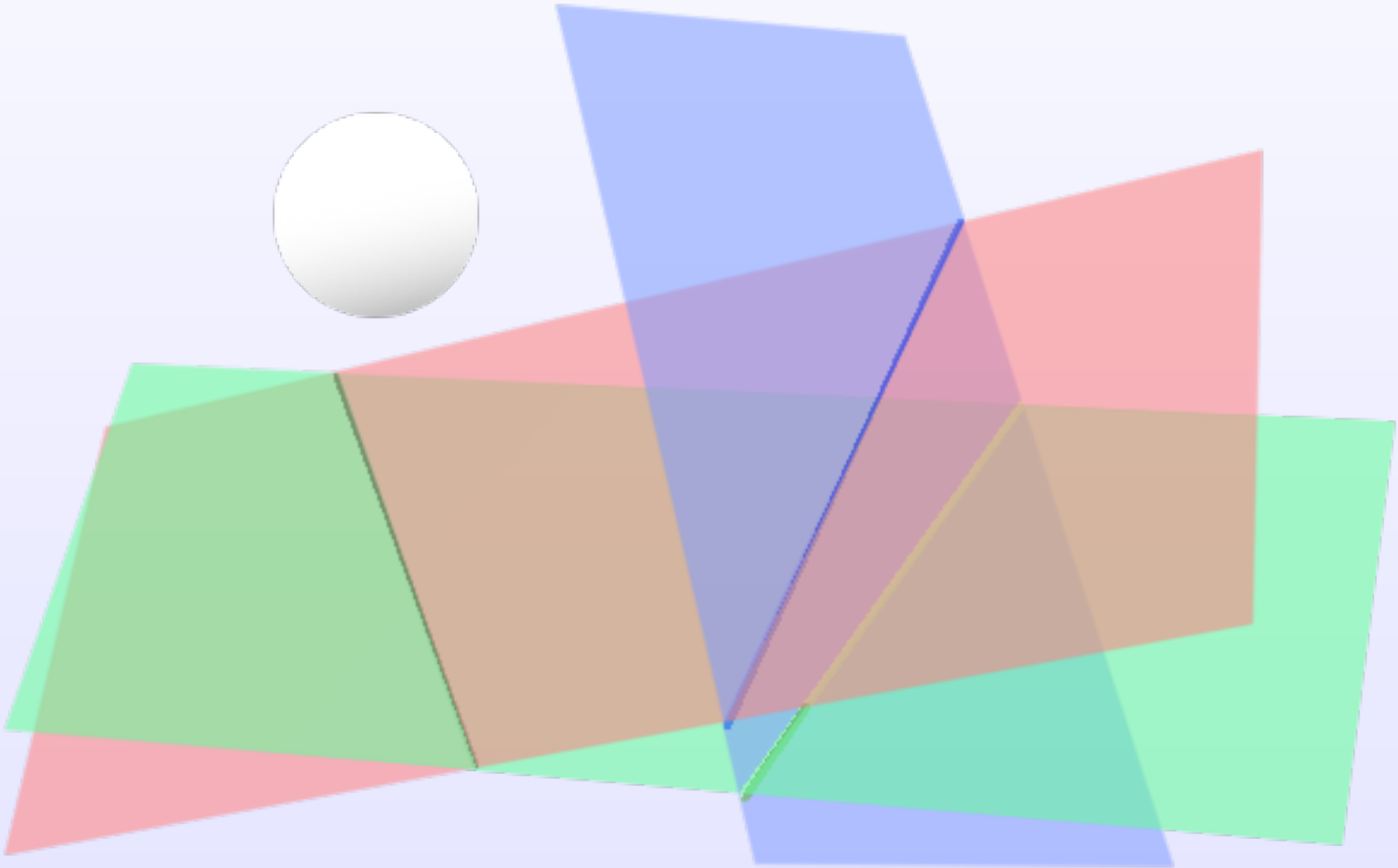
Kontextová skupina slova w s významem l obsahuje přesně ty jazykové situace, ve kterých má slovo w význam l.

$\mathcal{C}_h = \{c \in \mathcal{C} | \phi(\text{srážka}, c) = l\}$, kde l má význam „autonehoda“ a $\mathcal{C}_h$ je kontext „havárie“.

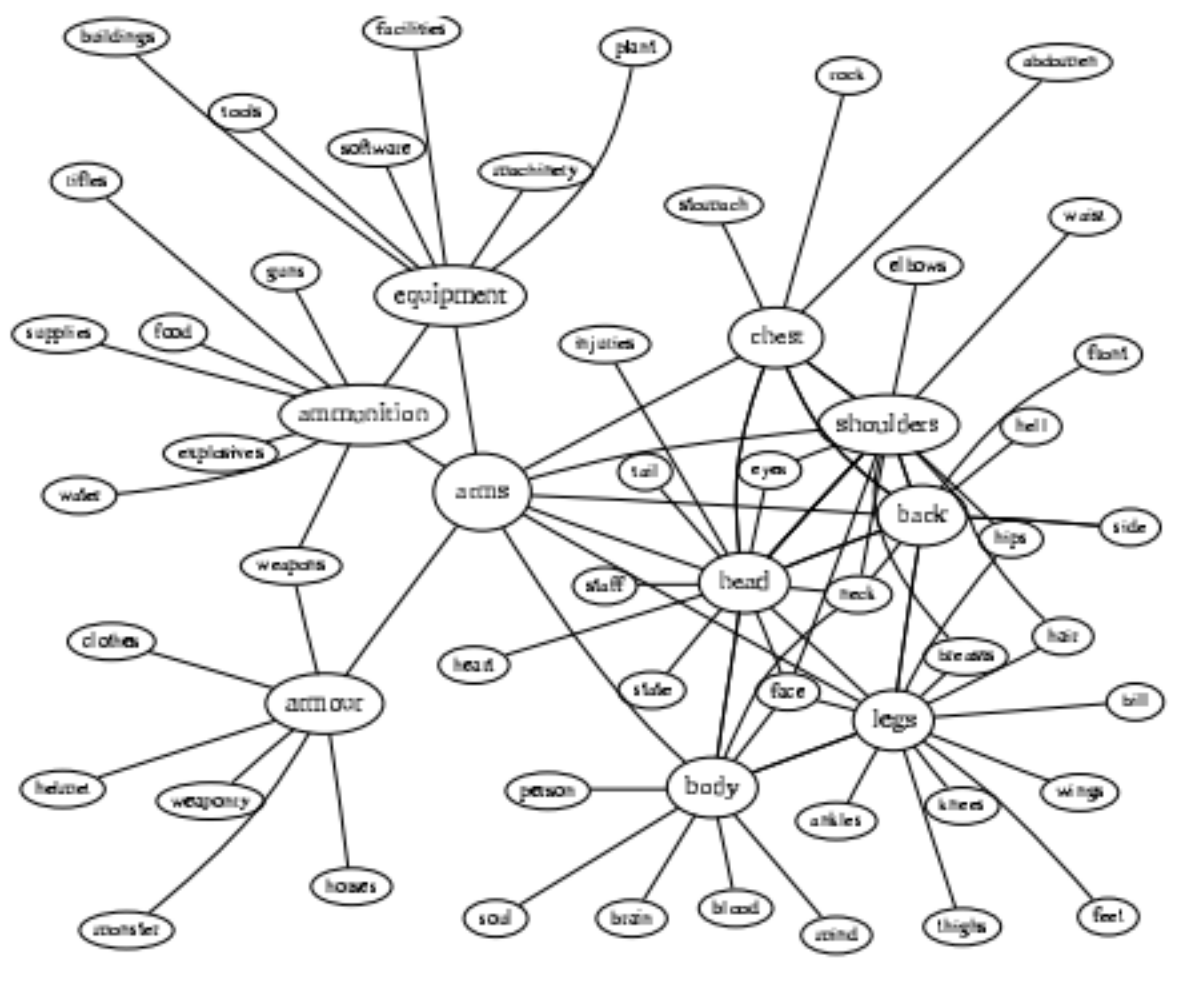
Kontext je inverzní zobrazení k přiřazení významu ϕ. Následkem toho, jsou v kontextu jen slova z okolí w, která lze použít k rozlišení významu.

Matematický model významu: kontextové skupiny

Umístit slovo do kontextu c_i je jako umístit kouli na nakloněnou rovinu:



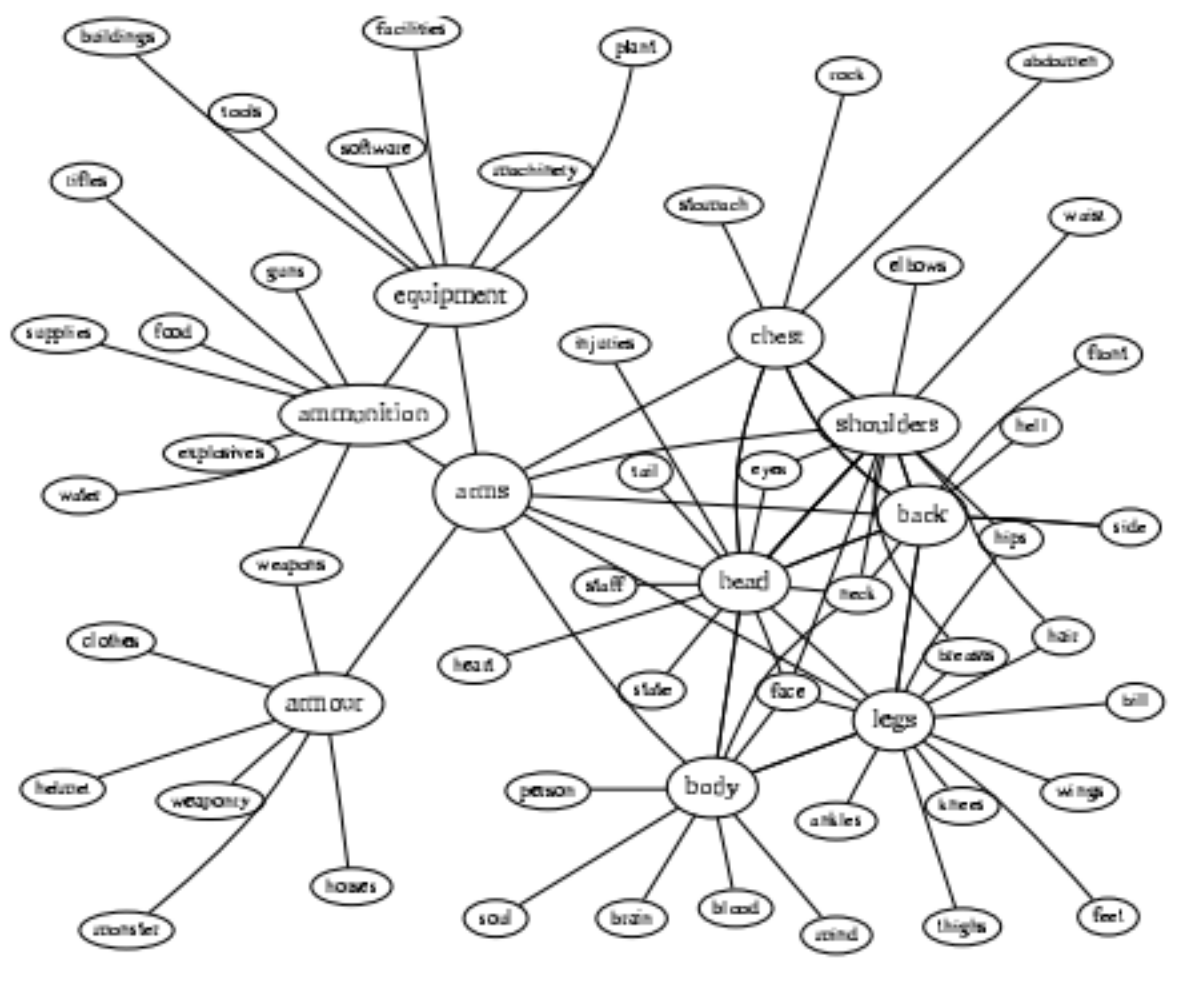
Matematický model významu: výsledek



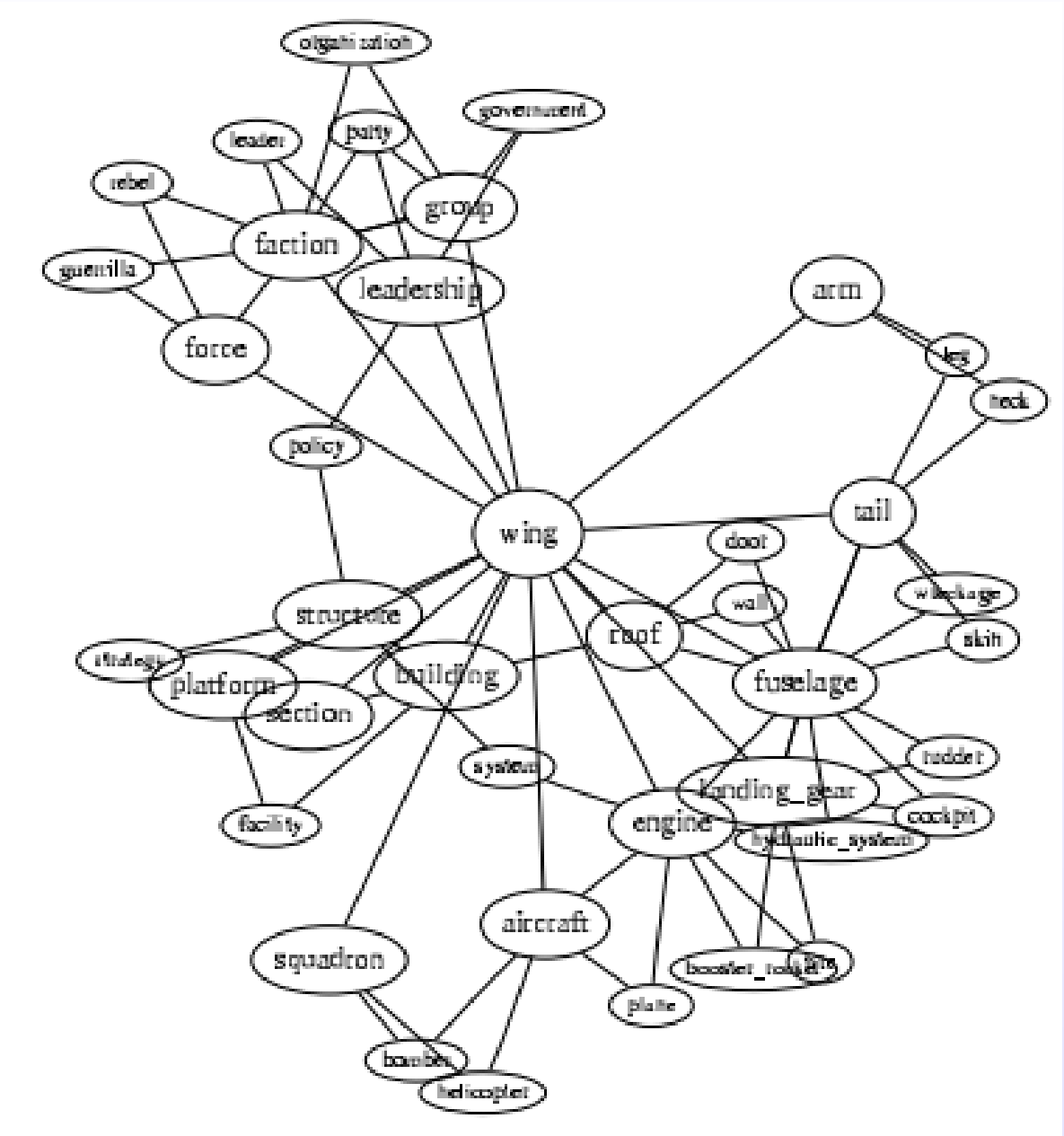
Matematický model významu: výsledek

- vzdálenost uzlu hraje roli (čím dál od sledovaného, tím méně ovlivňuje význam)
- propojenost uzlu s ostatními hraje roli (čím více propojení, tím větší víceznačnost)
- velikost uzlu hraje roli (čím menší, tím méně častý)

Matematický model významu: výsledek



Matematický model významu: výsledek



Hanks, P. (2000).
Do word meanings exist?
Computers and the Humanities, 34:205–215.
10.1023/A:1002471322828.

Hanks, P. (2010).
Elliptical arguments: a problem in relating meaning to use.
In Paquot, M. and Granger, S., editors, *eLexicography in the 21st century : New challenges, new applications. Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*, volume 7 of *Cahiers du Cental*, pages 109–124, Louvain-la-Neuve, Belgium. Université Catholique de Louvain, Presses universitaires de Louvain.

Král, R. (2006).
Word sense discrimination for czech.
In Sojka, P., Kopecek, I., and Pala, K., editors, *Text, Speech and Dialogue*, volume 2448 of *Lecture Notes in Computer Science*, pages 155–158. Springer Berlin / Heidelberg.

Schütze, H. (1998).
Automatic word sense discrimination.
Comput. Linguist., 24:97–123.

Widdows, D. (2003).
A mathematical model for context and word-meaning.
In *Proceedings of the 4th international and interdisciplinary conference on Modeling and using context*, CONTEXT'03, page 369–382, Berlin, Heidelberg. Springer-Verlag.