PDEV

Klasifikátor patternov

Vincent Kríž
ÚFAL, MFF UK | vincent.kriz@kamadu.eu

Obsah

- **Úvod**
- **Definovanie úlohy, popis vstupných údajov**
  - Klasifikácia patternov ako úloha strojového učenia
  - Skúmanie vstupných údajov
  - Skúmanie výstupných tried
- **Metódy a experimenty**
  - Baseline
  - Bag-of-words
  - Defaultná sada rysov
- **Záver**

Obsah

Corpus Pattern Analysis

- John Sinclair o WSD:
  - v autentickom použití jazyka je väčšina ambiguít riešiteľná pomocou kontextu
  - lexikálna jednotka by mala byť popísaná syntagmaticky i paradigmaticky
- Patrick Hanks: Corpus Pattern Analysis (**CPA**)
  - poloformálny lexikálny popis, ktorý konzistentne implementuje Sinclairov koncept zachytávania významu vo vzoroch použitia v jazyku namiesto lexikálnych jednotiek, ktoré sú používané v tradičnej lexikografii

Pattern Dictionary of English Verbs

- Pattern Dictionary of English Verbs (PDEV)
  - zachycuje ”normálne” použitia daného slovesa zotriedené do vzorov (**patternov**)
  - pattern pozostáva z lematizovaného slovesa a relevantných **kolokácií**
  - kolokácie sú klasifikovaná pomocou **sémantických typov**, **sémantických rolí** a **lexikálnych množín**
  - každá propozícia je **parafrázovaná** pomocou vety, v ktorej sú označené všetky relevantné kolokácie
  - parafráza predstavuje **implikatúru** alebo **významový potenciál** aktivovaný príslušným patternom

Pattern Dictionary of English Verbs

- Príklad:

2 **[[Human 1 = Student | Disciple]] follow [[Human 2 = Teacher]]**

[[Human 1 = Student | Disciple]] studies and is influenced by or tries to practice the teachings of [[Human 2 = Teacher]]

3 **[[Human | Institution]] follow [[Command | Document | Plan]]**

[[Human | Institution]] acts in accordance with [[Command | Plan]] (expressed in [[Document]])

4 **[[Event 1]] follow ([[Event 2]])**

[[Event 1]] happens after and typically as a consequence of [[Event 2]]

9 **[[Artifact | Proposition]] answer {need | purpose}**

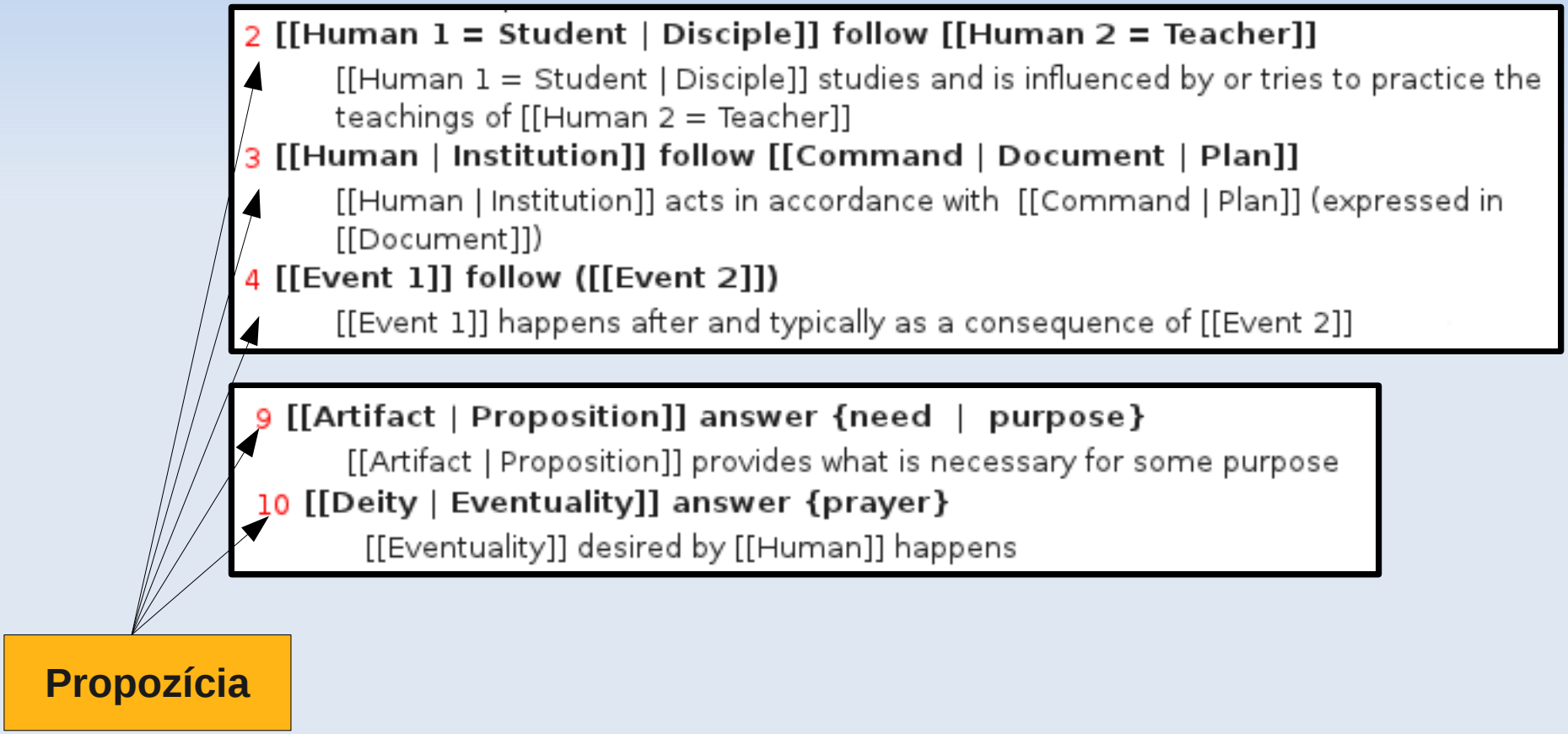
[[Artifact | Proposition]] provides what is necessary for some purpose

10 **[[Deity | Eventuality]] answer {prayer}**

[[Eventuality]] desired by [[Human]] happens

Pattern Dictionary of English Verbs

- Príklad:



Pattern Dictionary of English Verbs

- Príklad:

2 **[[Human 1 = Student | Disciple]] follow [[Human 2 = Teacher]]**

[[Human 1 = Student | Disciple]] studies and is influenced by or tries to practice the teachings of [[Human 2 = Teacher]]

3 **[[Human | Institution]] follow [[Command | Document | Plan]]**

[[Human | Institution]] acts in accordance with [[Command | Plan]] (expressed in [[Document]])

4 **[[Event 1]] follow ([[Event 2]])**

[[Event 1]] happens after and typically as a consequence of [[Event 2]]

9 **[[Artifact | Proposition]] answer {need | purpose}**

[[Artifact | Proposition]] provides what is necessary for some purpose

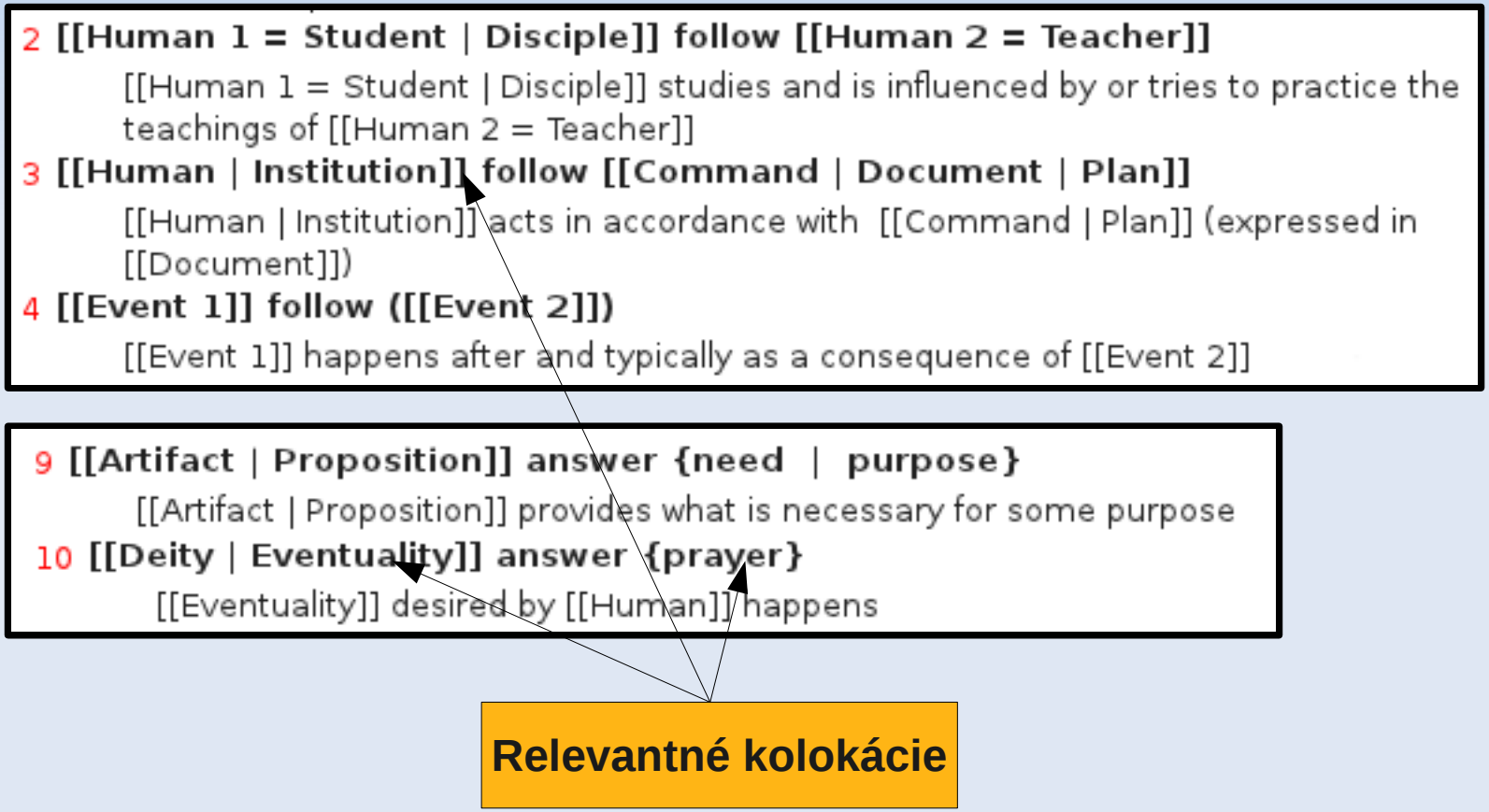
10 **[[Deity | Eventuality]] answer {prayer}**

[[Eventuality]] desired by [[Human]] happens

Lematizované sloveso

Pattern Dictionary of English Verbs

- Príklad:



Pattern Dictionary of English Verbs

- Príklad:

2 **[[Human 1 = Student | Disciple]] follow [[Human 2 = Teacher]]**

[[Human 1 = Student | Disciple]] studies and is influenced by or tries to practice the teachings of [[Human 2 = Teacher]]

3 **[[Human | Institution]] follow [[Command | Document | Plan]]**

[[Human | Institution]] acts in accordance with [[Command | Plan]] (expressed in [[Document]])

4 **[[Event 1]] follow ([[Event 2]])**

[[Event 1]] happens after and typically as a consequence of [[Event 2]]

9 **[[Artifact | Proposition]] answer {need | purpose}**

[[Artifact | Proposition]] provides what is necessary for some purpose

10 **[[Deity | Eventuality]] answer {prayer}**

[[Eventuality]] desired by [[Human]] happens

Parafráza

Pattern Dictionary of English Verbs

- Príklad:

2 **[[Human 1 = Student | Disciple]] follow [[Human 2 = Teacher]]**

[[Human 1 = Student | Disciple]] studies and is influenced by or tries to practice the teachings of [[Human 2 = Teacher]]

3 **[[Human | Institution]] follow [[Command | Document | Plan]]**

[[Human | Institution]] acts in accordance with [[Command | Plan]] (expressed in [[Document]])

4 **[[Event 1]] follow ([[Event 2]])**

[[Event 1]] happens after and typically as a consequence of [[Event 2]]

9 **[[Artifact | Proposition]] answer {need | purpose}**

[[Artifact | Proposition]] provides what is necessary for some purpose

10 **[[Deity | Eventuality]] answer {prayer}**

[[Eventuality]] desired by [[Human]] happens

Sémantická rola

Pattern Dictionary of English Verbs

- Príklad:

2 **[[Human 1 = Student | Disciple]] follow [[Human 2 = Teacher]]**

[[Human 1 = Student | Disciple]] studies and is influenced by or tries to practice the teachings of [[Human 2 = Teacher]]

3 **[[Human | Institution]] follow [[Command | Document | Plan]]**

[[Human | Institution]] acts in accordance with [[Command | Plan]] (expressed in [[Document]])

4 **[[Event 1]] follow ([[Event 2]])**

[[Event 1]] happens after and typically as a consequence of [[Event 2]]

9 **[[Artifact | Proposition]] answer {need | purpose}**

[[Artifact | Proposition]] provides what is necessary for some purpose

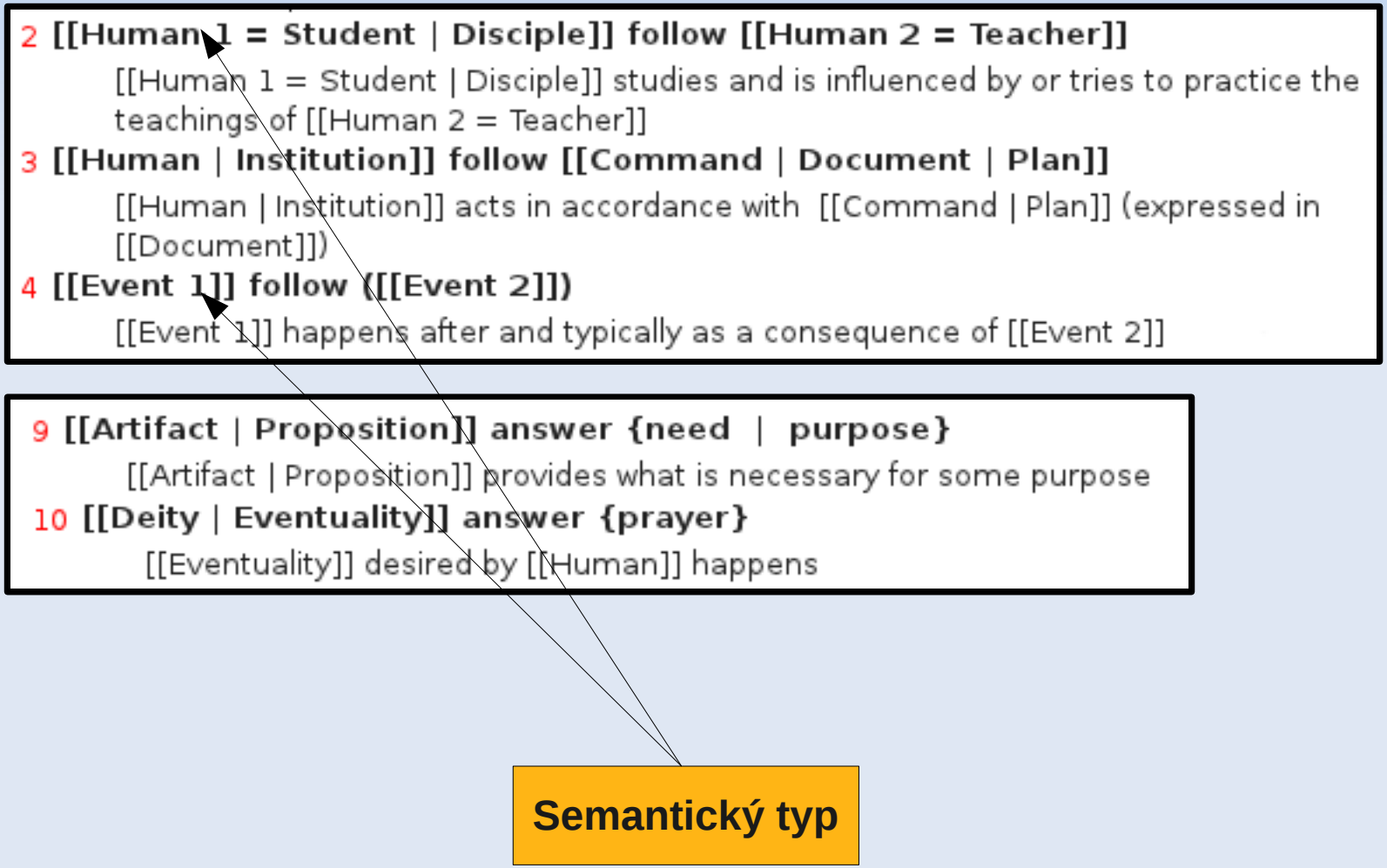
10 **[[Deity | Eventuality]] answer {prayer}**

[[Eventuality]] desired by [[Human]] happens

Lexikálna množina

Pattern Dictionary of English Verbs

- Príklad:



Obsah

- **Úvod**
- **Definovanie úlohy, popis vstupných údajov**
  - Klasifikácia patternov ako úloha strojového učenia
  - Skúmanie vstupných údajov
  - Skúmanie výstupných tried
- **Metódy a experimenty**
  - Baseline
  - Bag-of-words
  - Defaultná sada rysov
- **Záver**

Zadanie úlohy

- **Motivácia**
  - CPA považujeme za užitočnú a jasne pochopiteľnú metódu
  - doteraz ale chýba dôkaz, že CPA je vhodná aj na strojové spracovanie jazyka
- **Vstup**
  - 30 anglických slovies
  - cca 300 manuálne anotovaných viet pre každé sloveso
- **Výstup**
  - aplikácia, ktorá zadanej vete priradí pattern

Dostupné údaje

- **30 anglických slovies**
  - cca 300 manuálne označkovaných viet
  - líšia sa pokrytím v korpuse
  - líšia sa distribúciou patternov
  - líšia sa medzianotátorskou zhodou
- **10 slovies pre podrobné experimenty**
  - reprezentanti z celého spektra charakteristík
    - slovesá s malou aj veľkou frekvenciou
    - slovesá s malým aj veľkým množstvom patternov
    - slovesá s malou aj veľkou perplexitou

Metodika riešenia úlohy

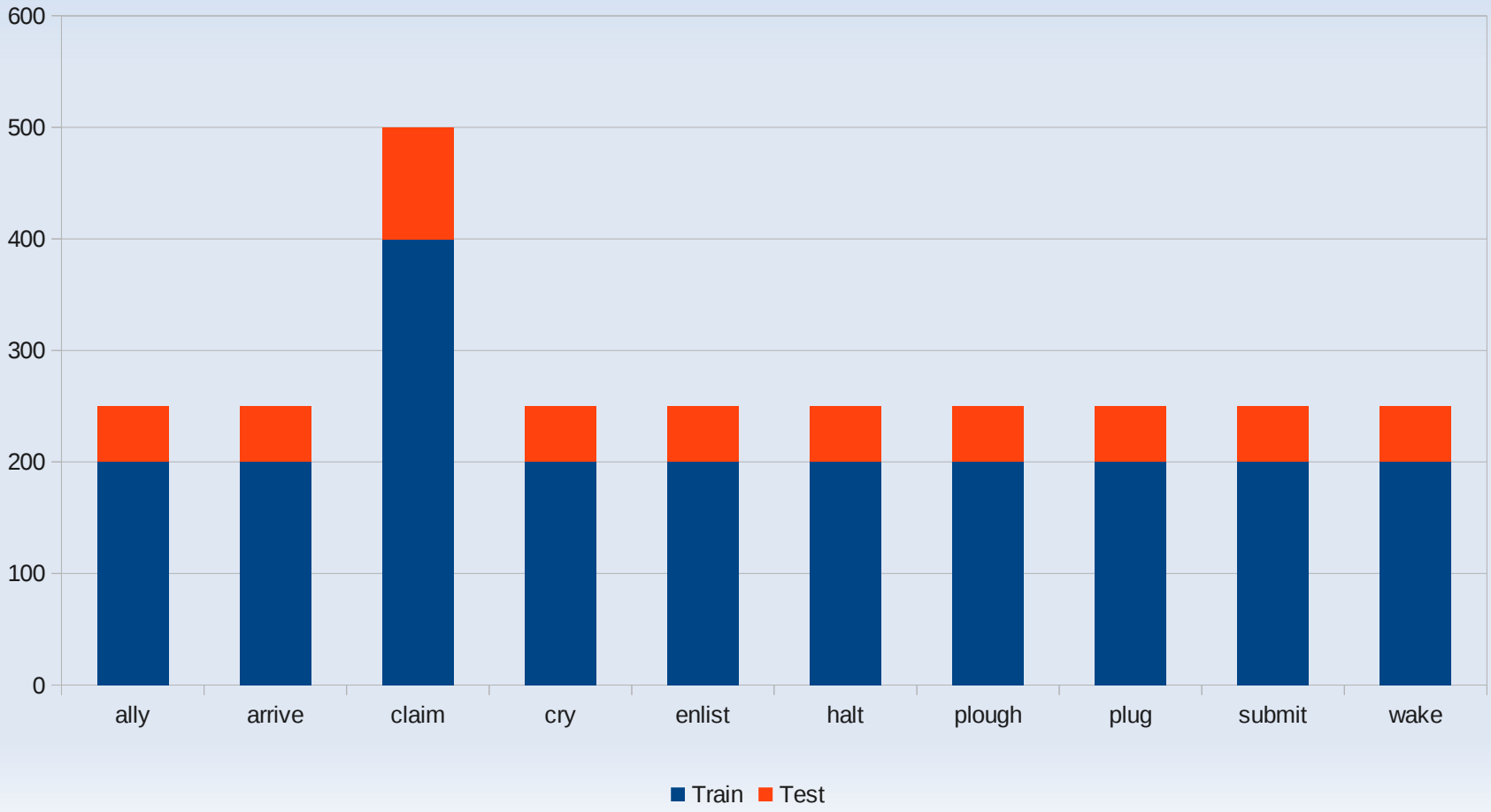
- **10** *pilotných* slovies **odladíme ručne**
  - vymyslíme vhodnú reprezentáciu viet
  - vyladíme parametre algoritmov strojového učenia
- **20** slovies **spracujeme automaticky**
  - určíme, ktorému z 10 pilotných slovies sa sloveso podobá najviac
  - podľa toho určíme vhodnú reprezentáciu viet
  - podľa toho nastavíme parametre algoritmov strojového učenia

Obsah

- **Úvod**
- **Definovanie úlohy, popis vstupných údajov**
  - Klasifikácia patternov ako úloha strojového učenia
  - Skúmanie vstupných údajov
  - Skúmanie výstupných tried
- **Metódy a experimenty**
  - Baseline
  - Bag-of-words
  - Defaultná sada rysov
- **Záver**

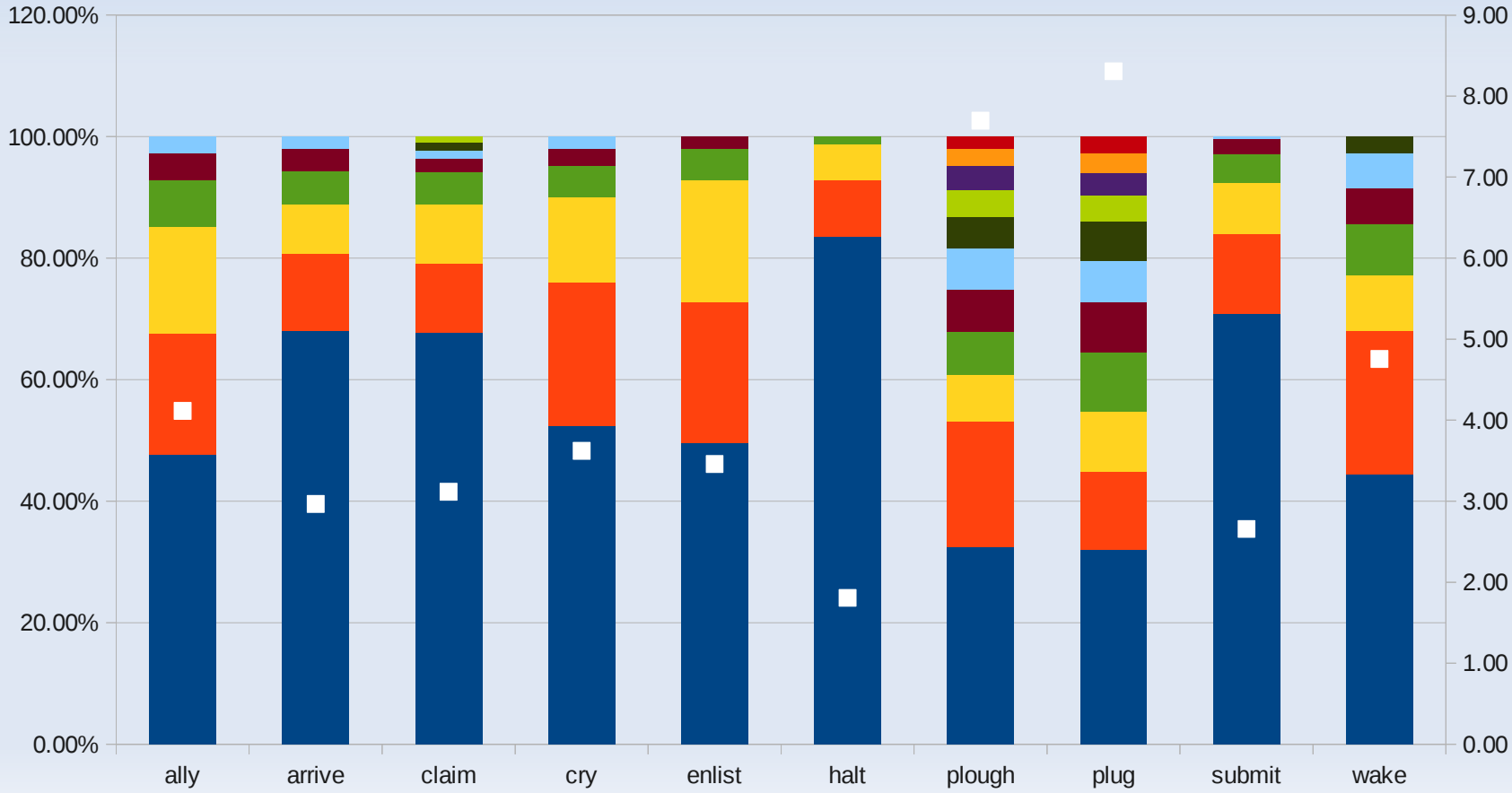
Čo vieme o vstupných údajoch

- **Počet dostupných údajov**



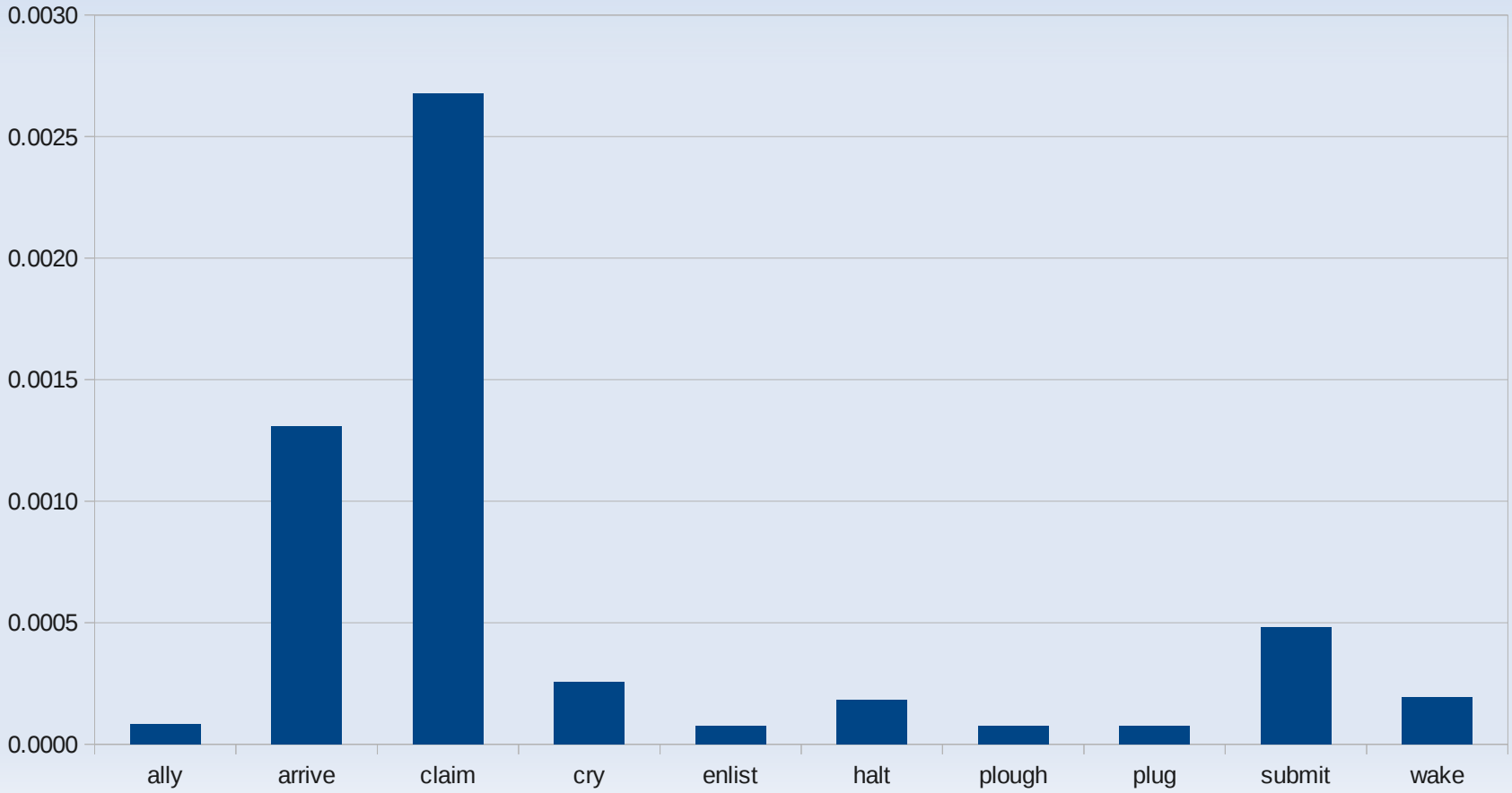
Čo vieme o vstupných údajoch

- **Distribúcia patternov, perplexita**



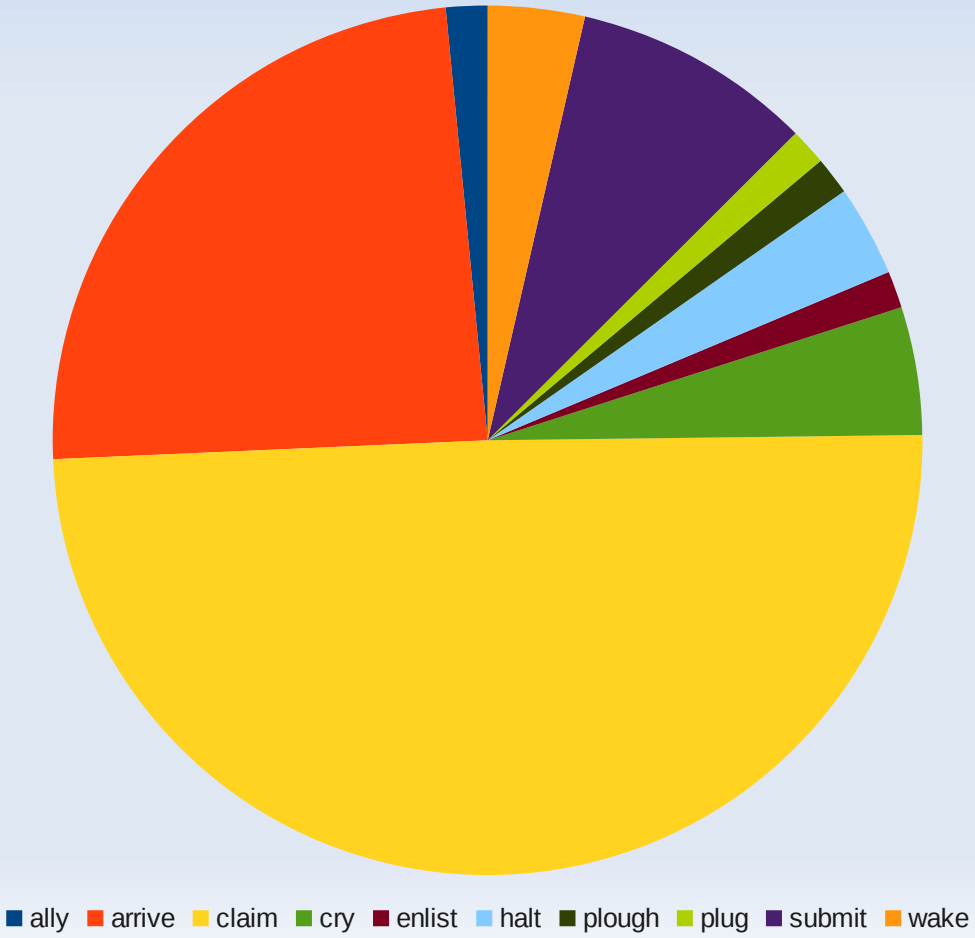
Čo vieme o vstupných údajoch

- **Pokrytie v BNC50**



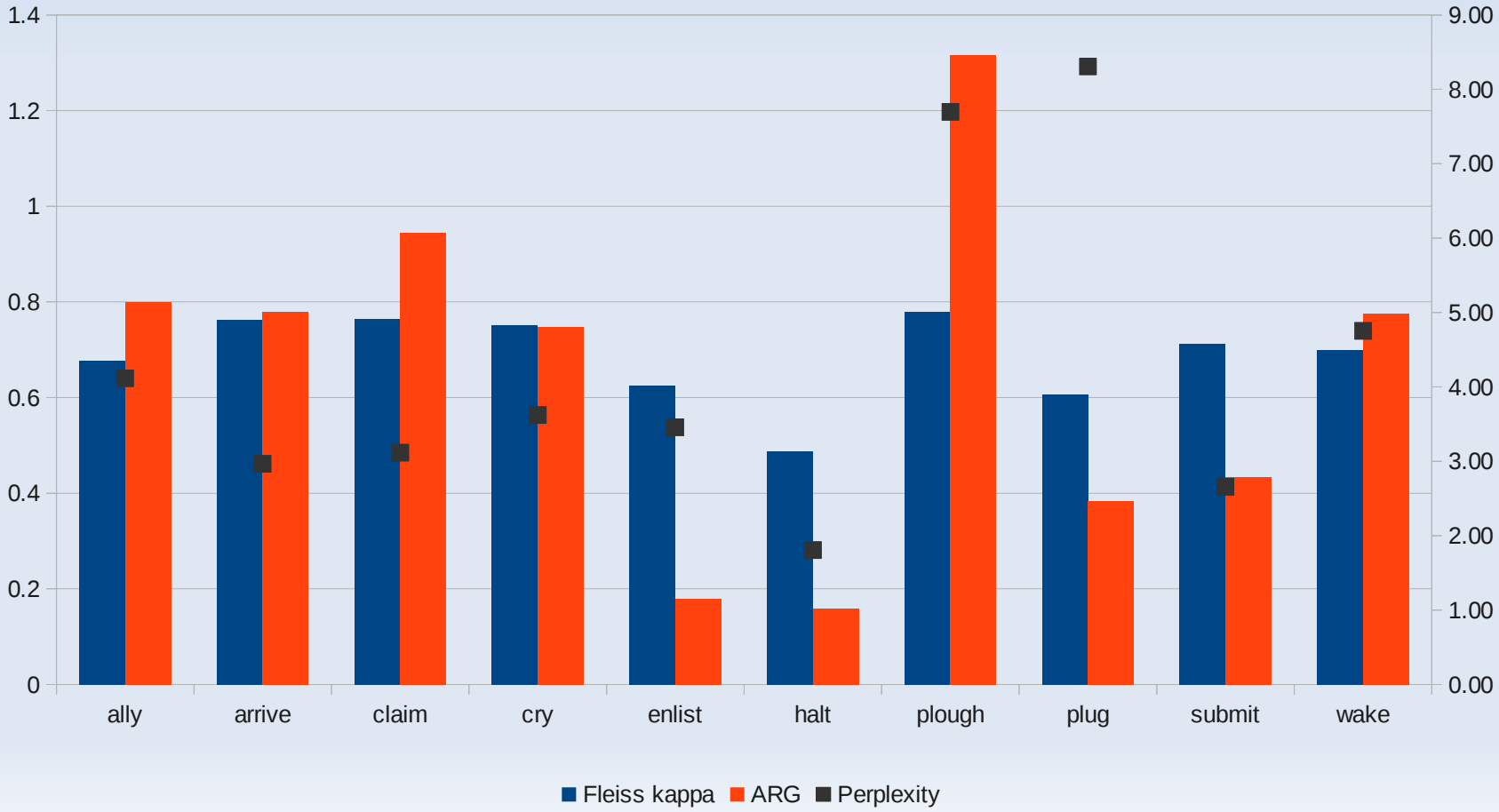
Čo vieme o vstupných údajoch

- **Pokrytie v BNC50**



Čo vieme o vstupných údajoch

- **Medzianotátorská zhoda**



Obsah

Výstupné triedy

- **Normálne použitia** and **exploatácie**
- Anomálny argument (.a)
  - [[Human | Vehicle | Animal]] arrive [NO OBJ] {at [[Location]]}
  - The plot had arrived at Beirut.
- Coercion (.c)
  - [[Human]] drink [[Beverage]]
  - She drank 8 glasses of spirits
- Figuratívne použitie (.f)
  - [[Human | Institution]] arrive [NO OBJ] {at [[Concept = Considered Opinion]]}
  - Nobody arrives at ICI board level without some steel in his character.

Výstupné triedy

- Anomálna syntax (.s)
  - [[Human 1 | Institution 1]] legally imposes a fine on, imprisons, or inflicts harm on [[Human 2 | Institution 2]] for [[Action]]
  - We punish too much—and in particular, we imprison too much
- Špeciálne značky **X** a **U**
- Chyby v taggovaní a ďalší šum (x)
  - **Ally** McCoist salutes Rangers' second goal
- Neklasifikovateľné (u)
  - Značka pre inštancie, ktorým nie je možné priradiť ani jeden z definovaných patternov

Výstupné triedy

- Patterny by mali byť triedy, ale
  - n patternov = $n + 4n + 1 + 1$ tried
- Nemáme dostatok inštancií pre každú triedu
  - potrebujeme aspoň 5-10
- **Aké sú možnosti?**
  - zlúžiť exploatácie s normálnymi použitiami
  - odstrániť triedy s malými frekvenciami
  - zlúčiť podobné patterny
- Ako ošetriť špeciálne triedy X a U?

Strojové učenie

- **Reprezentácia objektov z reálneho sveta**
  - vytvoríme množinu rysov, ktorými budeme objekty charakterizovať
  - **inštancie** budeme reprezentovať **vektormi rysov**

Kráľovná morí: <1500t, 5kW, 100m, 1976>
Victoria II: <1600t, 4kW, 105m, 1876>
Enez Euza: <100t, 1kW, 50m, 2006>

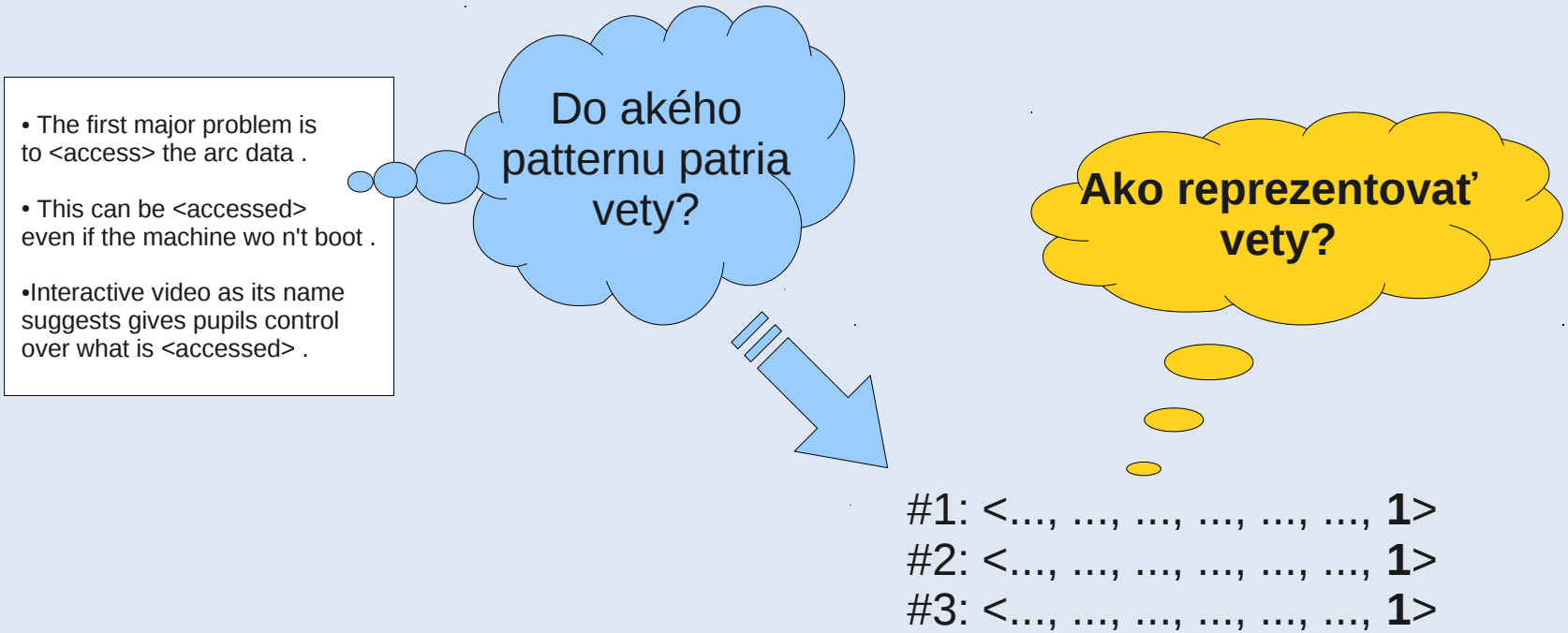
Strojové učenie

- **Klasifikačná úloha**
  - každej inštancii priradíme triedu z množiny **výstupných tried**
  - vedomosť, akú triedu priradzovať získame z inštancií, ktoré už túto triedu majú priradenú

Kráľovná morí: <1500t, 5kW, 100m, 1976, **T**>
Victoria II: <1600t, 4kW, 105m, 1876, **F**>
Enez Euza: <100t, 1kW, 50m, 2006, **F**>

Strojové učenie

- **Reprezentácia objektov z reálneho sveta**
  - vytvoríme množinu rysov, ktorými budeme objekty charakterizovať
  - **inštancie** budeme reprezentovať **vektormi rysov**

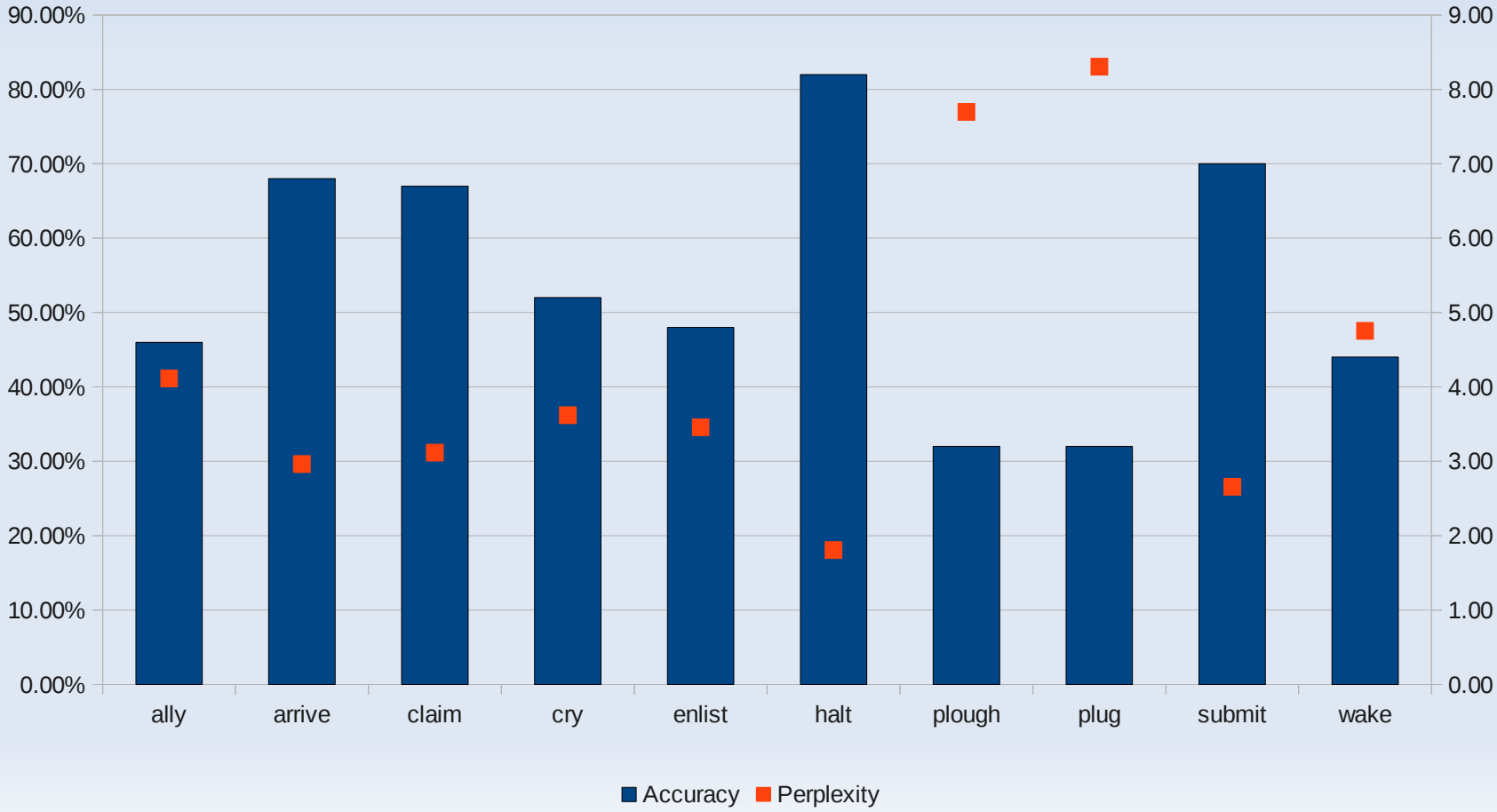


Obsah

- **Úvod**
- **Definovanie úlohy, popis vstupných údajov**
  - Klasifikácia patternov ako úloha strojového učenia
  - Skúmanie vstupných údajov
  - Skúmanie výstupných tried
- **Metódy a experimenty**
  - Baseline
  - Bag-of-words
  - Defaultná sada rysov
- **Záver**

Baseline

- Každej inštancii priradíme **najpravdepodobnejší** pattern

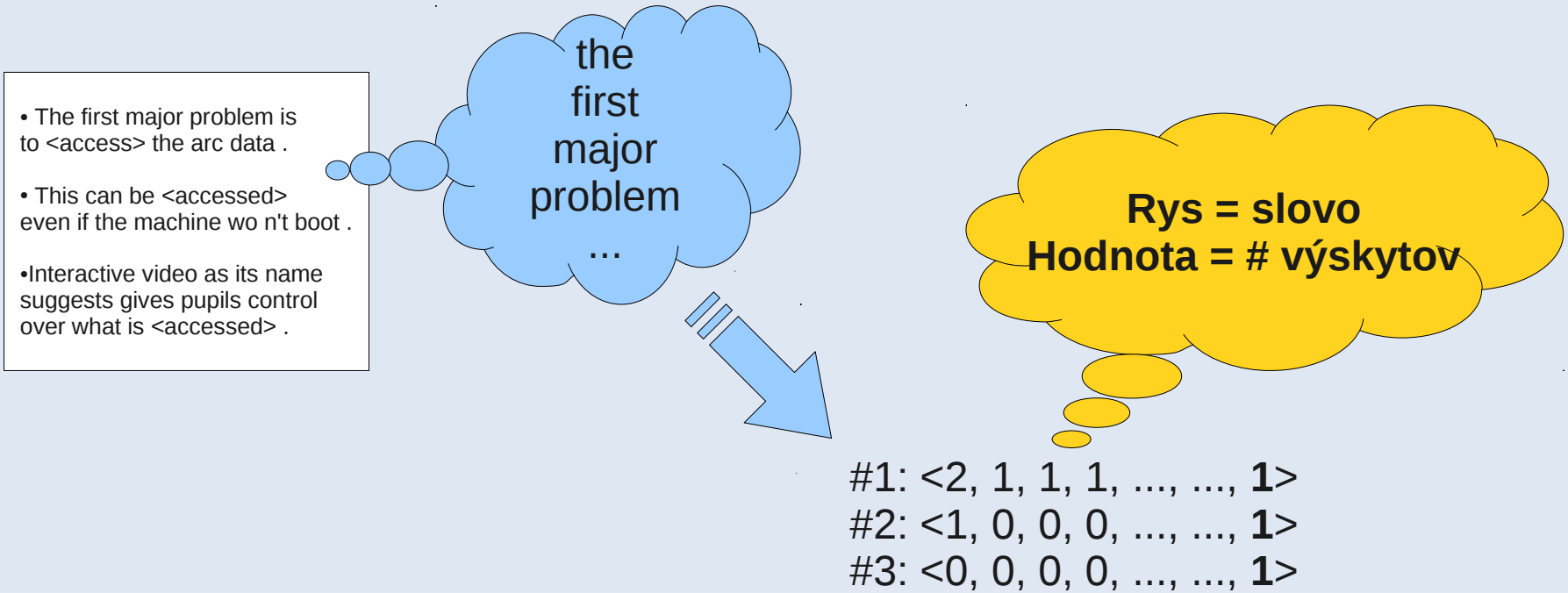


Obsah

- **Úvod**
- **Definovanie úlohy, popis vstupných údajov**
  - Klasifikácia patternov ako úloha strojového učenia
  - Skúmanie vstupných údajov
  - Skúmanie výstupných tried
- **Metódy a experimenty**
  - Baseline
  - Bag-of-words
  - Defaultná sada rysov
- **Záver**

Bag-of-words

- **Je znalosť slov okolo cieľového slovesa dostatočná na klasifikáciu patternov?**
  - množina rysov bude tvorená slovami
    - rys kóduje informáciu, koľkokrát sa dané slovo vyskytuje v klasifikovanej vete

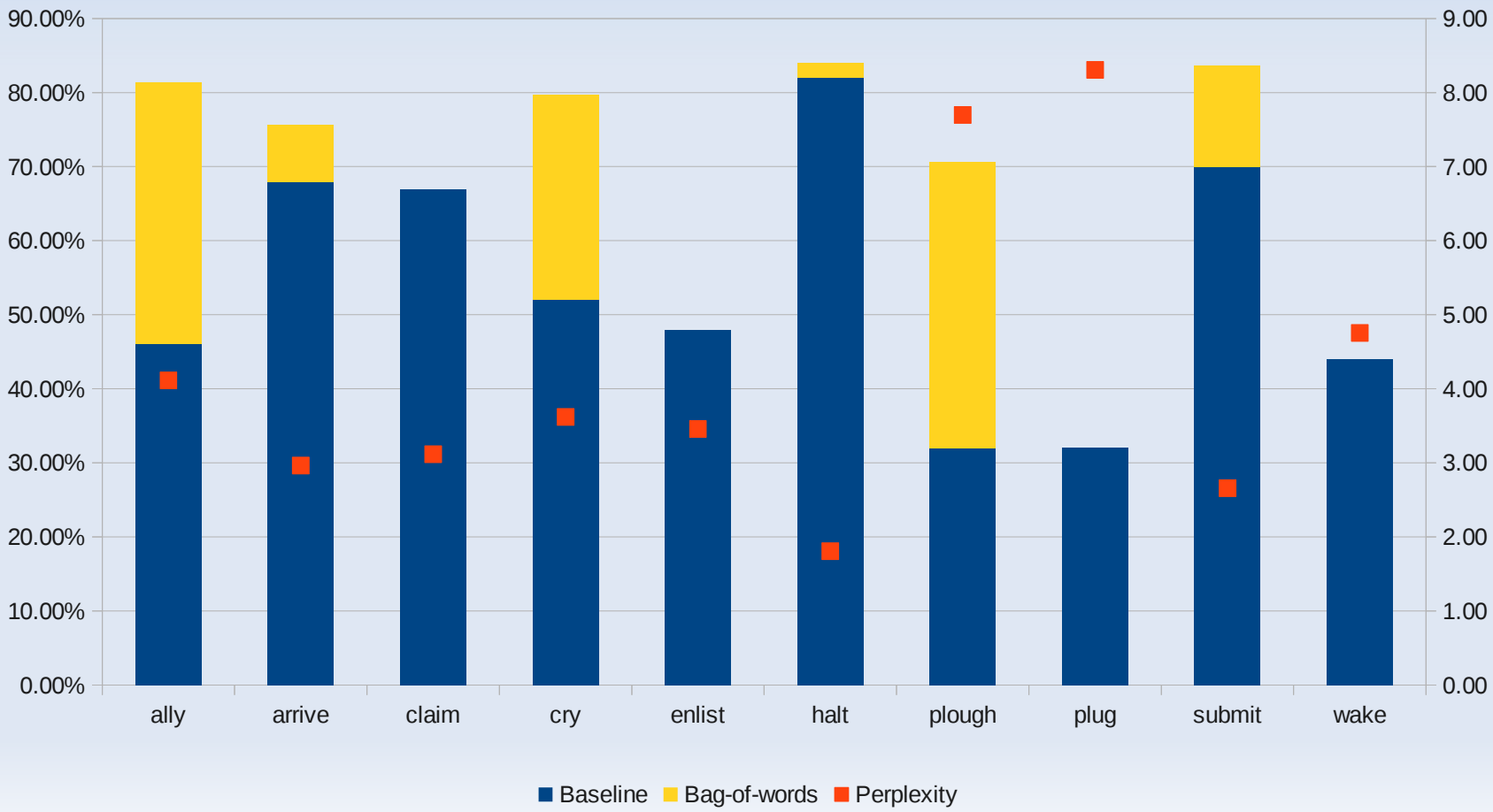


Bag-of-words

- Aké slová použiť v množine rysov?
  - najčastejšie slová z korpusu
  - slová vyskytujúce sa v trénovacích údajoch
- Experimenty s DT, kNN, SVM, **Adaboost**
- 10-fold cross-validation
  - robustnosť modelu
  - konfidenčný interval pre mieru úspešnosti

Bag-of-words

- **Výsledky:**



Obsah

- **Úvod**
- **Definovanie úlohy, popis vstupných údajov**
  - Klasifikácia patternov ako úloha strojového učenia
  - Skúmanie vstupných údajov
  - Skúmanie výstupných tried
- **Metódy a experimenty**
  - Baseline
  - Bag-of-words
  - Defaultná sada rysov
- **Záver**

Ďalšie prístupy

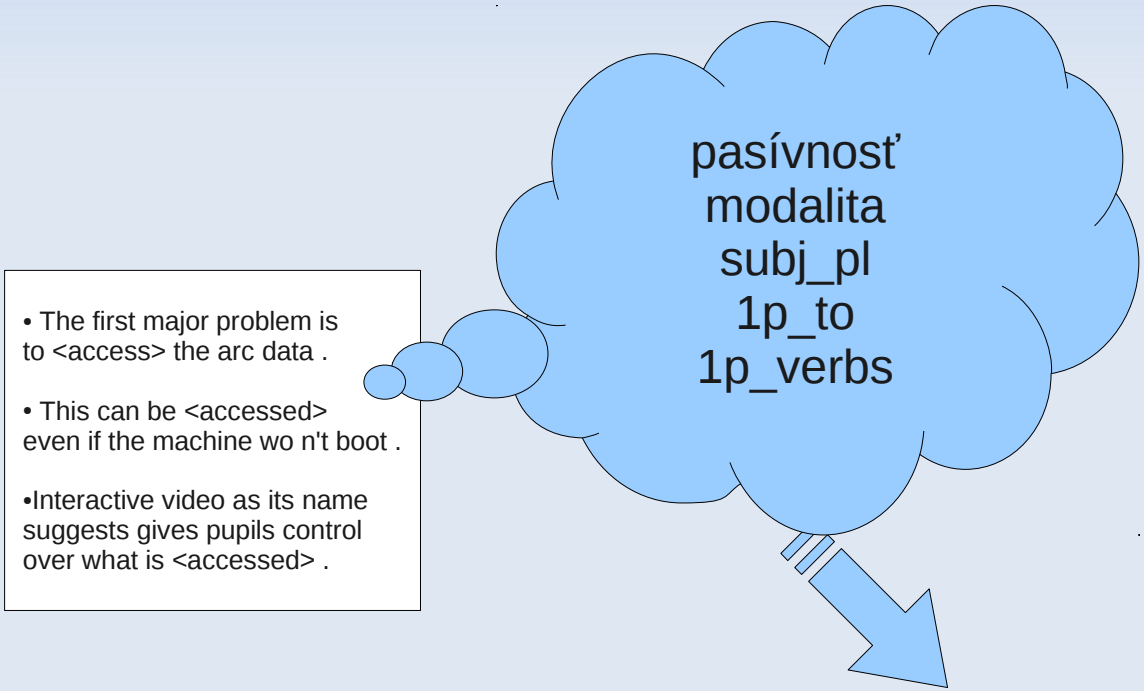
- Bag-of-words využíva *len* slová
- Môžme použiť **syntax** a **sémantiku**
- **Aké nástroje môžeme použiť?**
  - Tagging (Stanford POS tagger)
  - Parsing (Stanford parser)
  - Name entity recognition (Stanford NER)
  - WordNet
  - Populácie sémantickych typov
  - Triviálne pravidlá
    - napr. zámená he, she sú typ Human

Defaultná sada rysov

- Martin Holub a Silvie Cinková vytvotili *defaultnú* množinu rysov
- Použité nástroje:
  - Tagging
  - Parsing
  - Name entity recognition
  - WordNet
- Experimenty s DT, kNN, SVM, Adaboost
- 10-fold cross-validation

Defaultná sada rysov

- **Morfologicko-syntaktické rysy**
  - binárne rysy budú charakterizovať sloveso, syntaktické členy a kontext



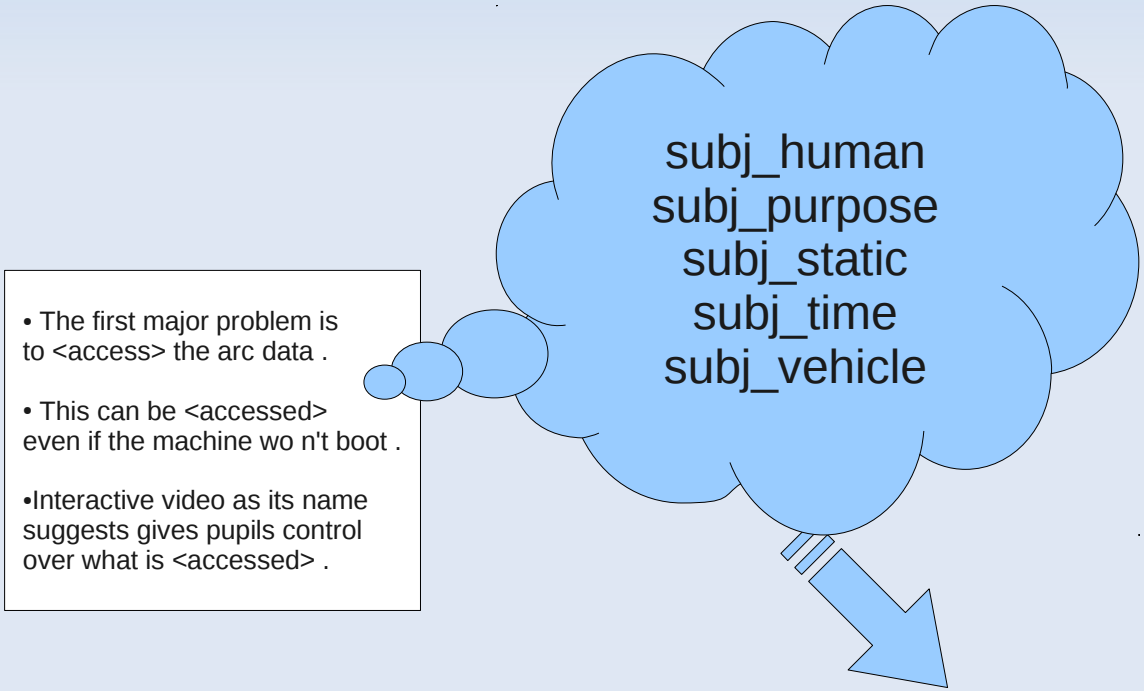
#1: <0, 0, 0, 1, 0, ..., **1**>
#2: <1, 1, 0, 0, 0, ..., **1**>
#3: <1, 0, 0, 0, 1, ..., **1**>

Defaultná sada rysov

- **Morfologicko-syntaktické rysy**
  - Charaketistika cieľového slovesa
    - napr. pasívnosť, modalita, negácia, čas
  - Charakteristika najbližšieho kontextu slovesa (3 slová pred slovesom a 3 slová za slovesom)
    - zaradenie slova do jednej z definovaných skupín (napr. substantíva, adjektíva, slovesá, modálne slovesá, príslovky, ...)
  - Charakterizácia syntakticky závislých slov
    - existencia subjektu, objektu, adverbiálov

Defaultná sada rysov

- **Sémantické rysy**
  - binárne rysy budú charakterizovať sémantický typ podmetu, predmetu a kontextu



#1: <0, 0, 0, 0, 0, ..., **1**>
#2: <0, 0, 0, 0, 0, ..., **1**>
#3: <0, 0, 0, 0, 0, ..., **1**>

Defaultná sada rysov

- **Sémantické rysy**
  - Zaradenie subjektu, objektu a dvoch najbližších substantív do jednej z 50 sémantických tried
  - 50 sémantických tried tvorí **Vossenovu ontológiu**, ktorá je vrcholom v hyperonymickej hierarchii synsetov vo WordNete
  - Ontológia bola použitá v práci p. Semeckého (nad českým WordNetom)
  - Nie je jasné mapovanie synsetov do ontológie :-(
- Vytvoríme priamo **mapovanie do Sémantických typov PDEV**

Iné spôsoby extrakcie sémantických rysov

- Hľadanie sémantických typov vo WordNete
- Aplikácia wn v Debiane

wn *party* -hypen

```
Sense 1

party, political party
        => organization, organisation
            => social group
                => group, grouping
                    => abstraction, abstract entity
                        => entity
```

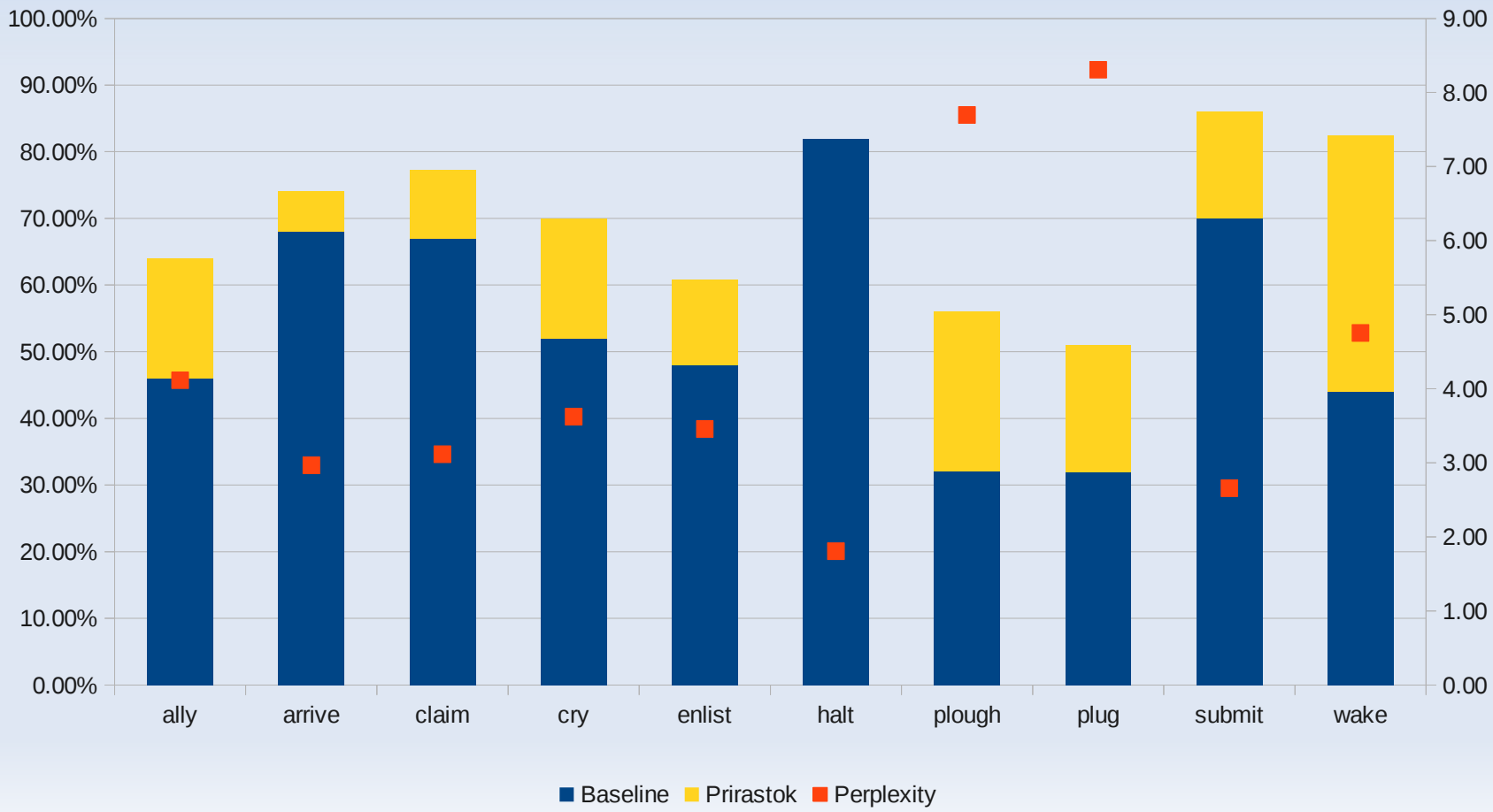
party je typ Organization

party je typ Institution

party je typ Entity

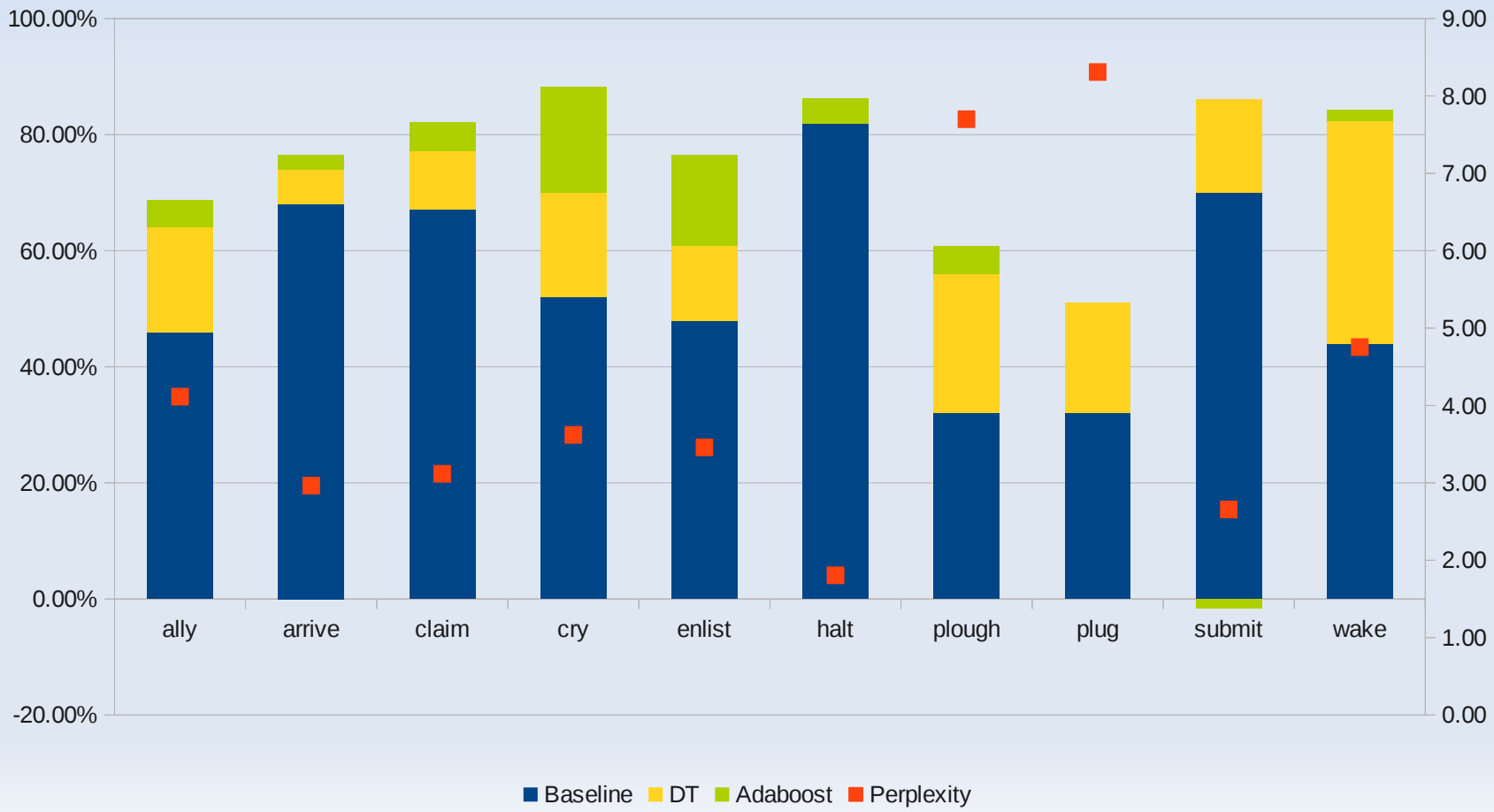
Výsledky experimentu

- **Rozhodovacie stromy (DT):**



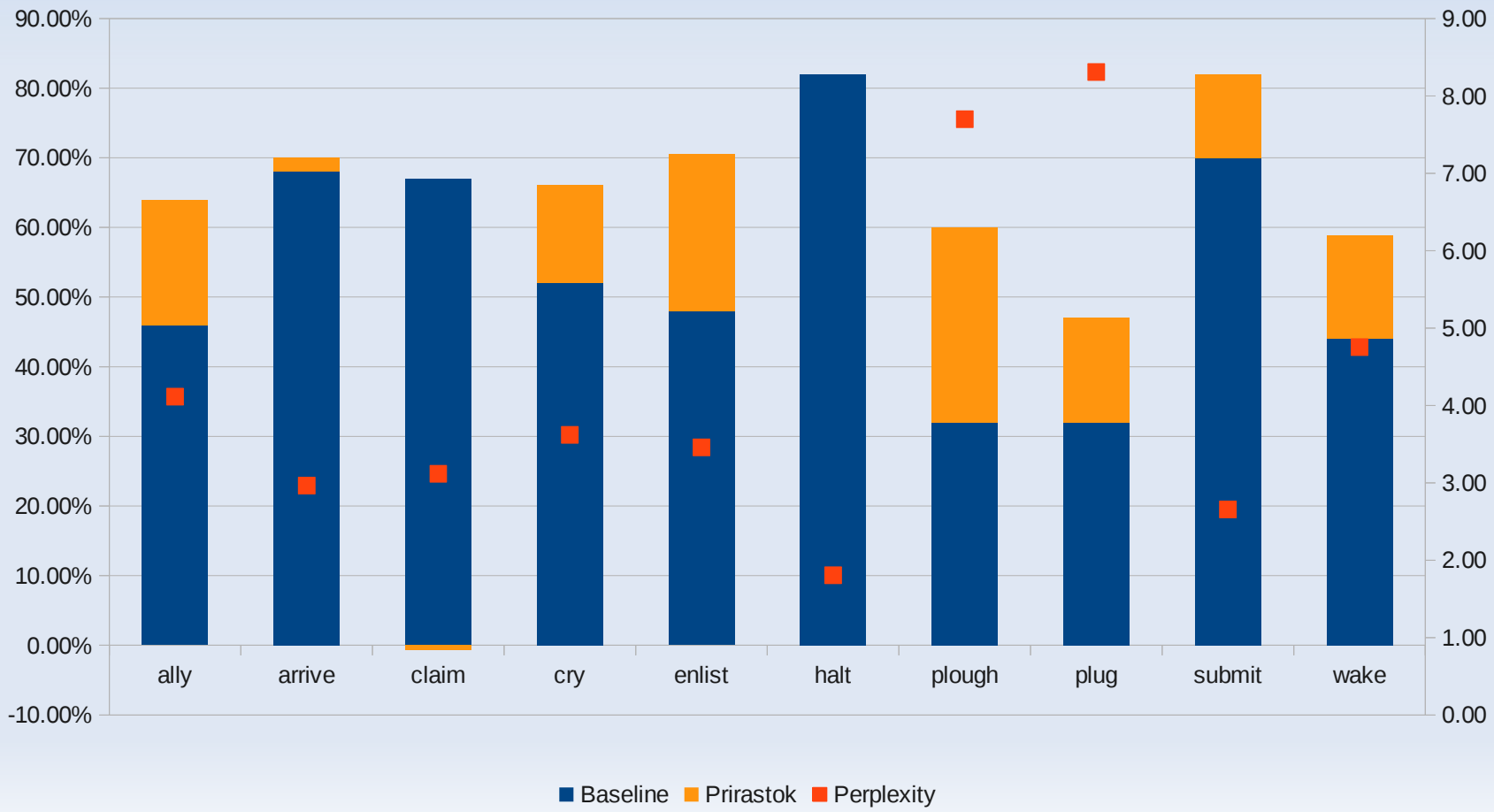
Výsledky experimentu

- **Adaboost.M1:**



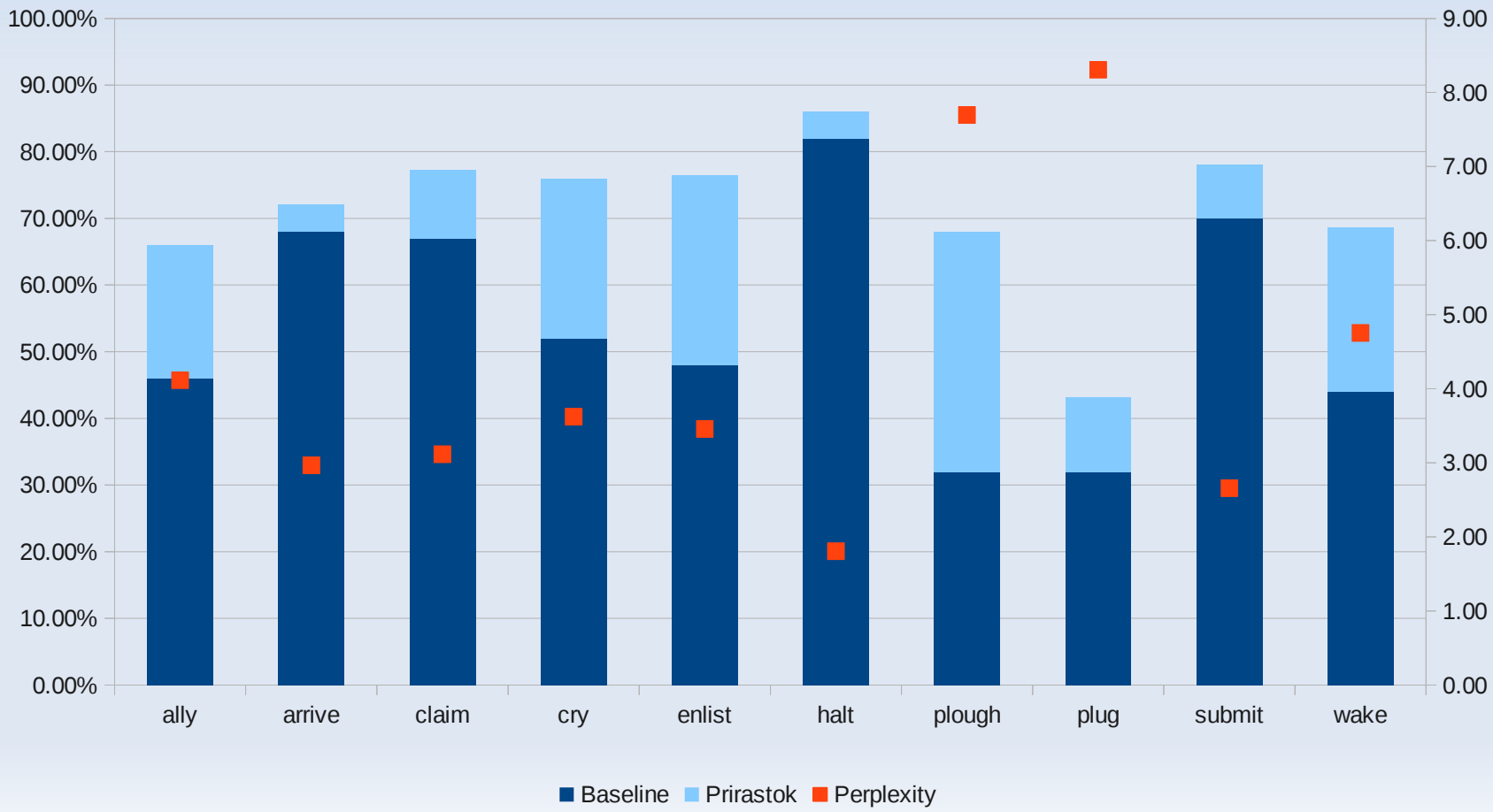
Výsledky experimentu

- **Najbližší susedia (kNN):**



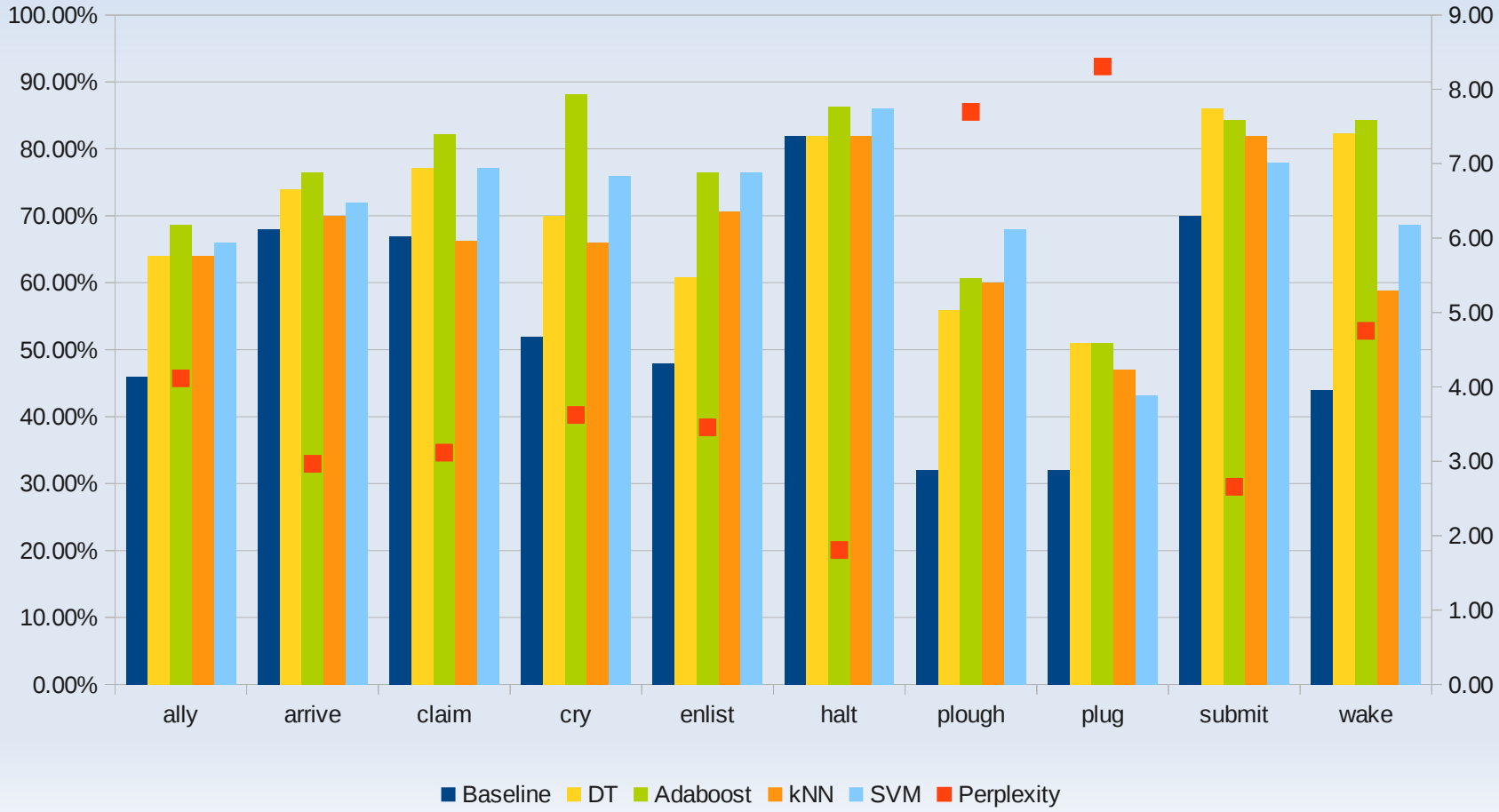
Výsledky experimentu

- **Podporné vektory (SVM):**



Výsledky experimentu

- **Súhrné výsledky:**



Obsah

- **Úvod**
- **Definovanie úlohy, popis vstupných údajov**
  - Klasifikácia patternov ako úloha strojového učenia
  - Skúmanie vstupných údajov
  - Skúmanie výstupných tried
- **Metódy a experimenty**
  - Baseline
  - Bag-of-words
  - Defaultná sada rysov
- **Záver**

Výhľady do budúcna :-)

- Vytvoriť lepšie **množiny rysov**
- Vytvoriť **mapovanie synsetov** WordNetu na Sémantické typy PDEV
  - vytvoriť populáciu Sémantických typov PDEV
- Zaradiť 20 slovies k podobným pilotným slovesám, evaluovať výsledky
- Vyhľadávať a používať **nové lexikálne zdroje**

Ďakujem za pozornosť :-)