Kapitola 23

Státnice I3: Návrh a vyhodnocování lingvistických experimentů

23.1 Úvod

- protože NLP používá ve velké míře stochastické metody, zaměříme se hlavně na experimenty testující účinnost těchto metod
- předtím, než je možné použít nějakou stochastickou metodu v praxi (a otesovat její účinnost), je nutné ji natrénovat na trénovacích datech
  - trénování závisí na dané metodě, ale většinou spočívá ve spočítání pravděpodobností použitých v metodě — často se odhadují pomocí relativních frekvencí získaných z trénovacích dat
- následně (pokud to daná metoda vyžaduje) je potřeba metodu přízpůsobit povaze dat (tzn. upravit její parametry, pokud existují), abychom maximalizovali její učinnost
  - parametry optimalizujeme na development datech
- následně je možné metodu otestovat pomocí vhodných metrik na testovacích datech

23.2 Příprava dat

- potřebujeme anotovaná data, u kterých ručně označíme správný výsledek experimentu — např. ručně přiřazené tagy slov pro tagging
- data je nutno rozdělit na 3 části
  - **training data** — největší, slouží k odhadnutí pravděpodobností; z velké části určují výsledek stochastické metody
  - **development data** — malá sada dat, která slouží k optimalizaci parametrů dané metody/modelu
  - **test data**
    - slouží pro ohodnocení kvality dané metody za použití vyhodnocovací metriky
    - nesmí být obsaženy v trénovacích a development datech, aby mohla být metoda objektivně ohodnocena
- pro nestochastické metody stačí pouze testovací data pro vyhodnocení

23.3 Standardní evaluační metriky

Evaluation

- test against evaluation test data – comparing the output of my parser to manually corrected data, done by someone else and in advance, independent of my algorithms
- rules:

1. should be automatic (if possible) – avoid subjective evaluation (but in e.g. SMT this is inevitable)
2. never tune the system using test data (use a small part of training data for this)
3. use standard metrics (if possible)

Hodnocení 1-1 metod

- pro každou vstupní jednotku vygeneruju jednu výstupní jednotku — např. tagging: každému slovu přiřadím tag
- **error rate**
- **accuracy**

Hodnocení 1-n metod

- délka vstupu a výstupu se může lišit — např. strojový překlad: výstupní věta může mít jinou délku než vstupní věta
- **precision**
- **recall**
- **f-measure**

Metriky strojového překladu

- BLEU
- NIST
- METEOR
  - upravená f-measure s důrazem na recall (precision:recall — 1:9)
  - párování slov na 3 úrovních: 1) slovní forma, 2) kořen slova, 3) WordNet synonymum
- PER (Position independent Error Rate), WER (Word Error Rate), TER (Translation Edit Rate), CDER

23.4 Typy evaluace podle úloh