Mallipõhine faktituletus tekstikorpustest

Eesti keeletehnoloogia projekt EKT22

1. aasta kokkuvõte

Sven Laur (swen@math.ut.ee)

Timo Petmanson (timo_p@ut.ee)

Ülesande püstitus

- Mida teha, kui tahame tekstidokumentidest või veebist eraldada informatsiooni.
- Ja me teame, mis tüüpi informatsiooniga on tegu:
  - **X** sündis aastal **Y**
  - Patsient kaebab, et **X**, **X**, **X** ja **X**.
  - Täna toimub **X** üritus **Y**.

Miks see on oluline?

- **Informatsioon on jõud!** Ja enamik informatsiooni on tänapäeval kirjeldatud vabatekstina.
- Palju tööd on tehtud ingliskeelsetest tekstidest faktide eraldamiseks: YAGO-NAGA, DBPedia, KnowItAll jne.

Kuidas tekstist fakte kätte saada?

- Võime kasutada programme nagu `grep, awk, sed` jt.

```
grep kaeb.* < medkorpus.txt
```

- Võime kasutada kitsendusgrammatikat ja defineerida omad mustrid

```
 SELECT (@NN>) (0 (P dem nom))(NOT -1
("kõik")) (*1 (S nom) ^ (@PRD) BARRIER
 CLB OR Öeldis OR Võrdlus OR (@PRD));
```

Meie lahendus

- Märgendame lausetes meid huvitavad **X** ja **Y** asukohad.
  - **Elva Kultuurikeskuses** algusega kell 13.00 algab **muusikakooli aktus**.
  - **Tallinnas Kuku Klubis** toimub ülehomme **maleõhtu**.
- “Õpime” märgendatud lausete abil hulga mustreid.
- Kasutame mustreid uute faktide otsimiseks märgendamata tekstides.

Märgendamine

Õppimine

- Esimese asjana eraldame märgendatud lauseosade põhjal spetsiifilised mustrid (iga märgendatud näite kohta üks)

*Meie president **Toomas Hendrik Ilves** ja **Barack Obama** kohtuvad järgmisel nädalal Kadrious.*

*Meie president **X** ja **Y** kohtuvad järgmisel nädalal Kadrious.*

Üldistamine konteksti vähendades

- Selleks, et õppida, peame suutma teadmisi üldistada.
  - Meie president **X** ja **Y** kohtuvad järgmisel nädalal Kadrious.
  - Meie president **X** ja **Y** kohtuvad järgmisel nädalal
  - …
  - **X** ja **Y**

Keeleliste atribuutide üldistamine

inessiiv: Kadriorus

sisekohakääne: Kadriorgu, Kadriorus, Kadriorust

kohakääne: Kadriorgu, Kadriorus, Kadriorust, Kadriorule, Kadriorul, Kadriorult

kõik käänded: Kadriorg, Kadrioru, …, Kadrioruga

Mustrite üldistamise algoritm

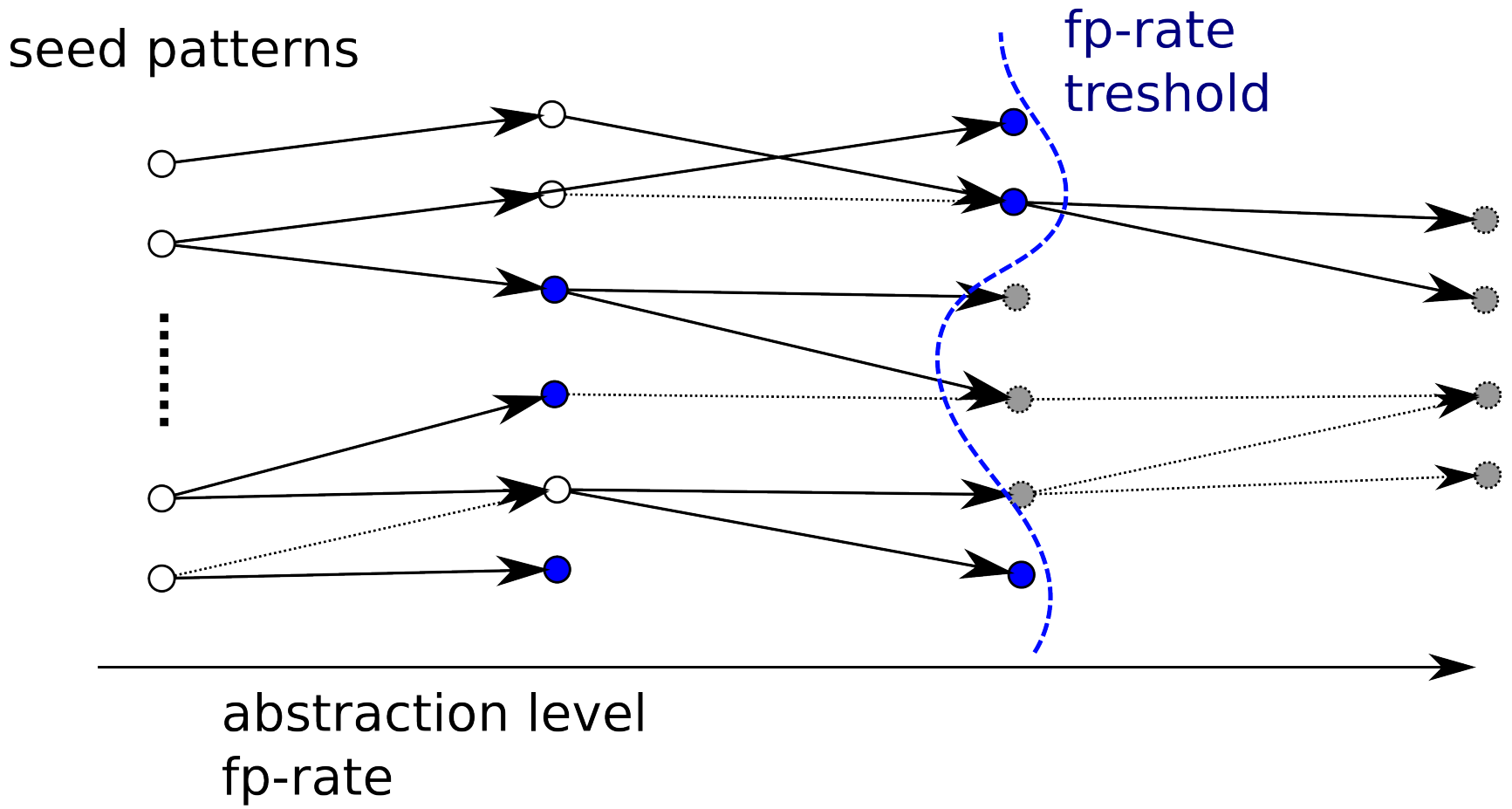
- Iga üldistussamm võib juurde tekitada valepositiivseid.
  - **X** ja **Y** kohtuvad
  - **X** ja **Y**

  Ilves ja **Obama** kohtuvad Kadriorus.

  Ain ja **Jüri** einestasid Rüütli tänaval.
- Määrame lubatud valepositiivsete osakaalu treeningkorpusel (*false positive rate*) ühe mustri jaoks. Vahetada punktid.

Mustrite otsimise algoritm

- Algoritmi käigus otsime maksimaalseid üldistusi, mis jäävad lubatud vigade tasemeni.

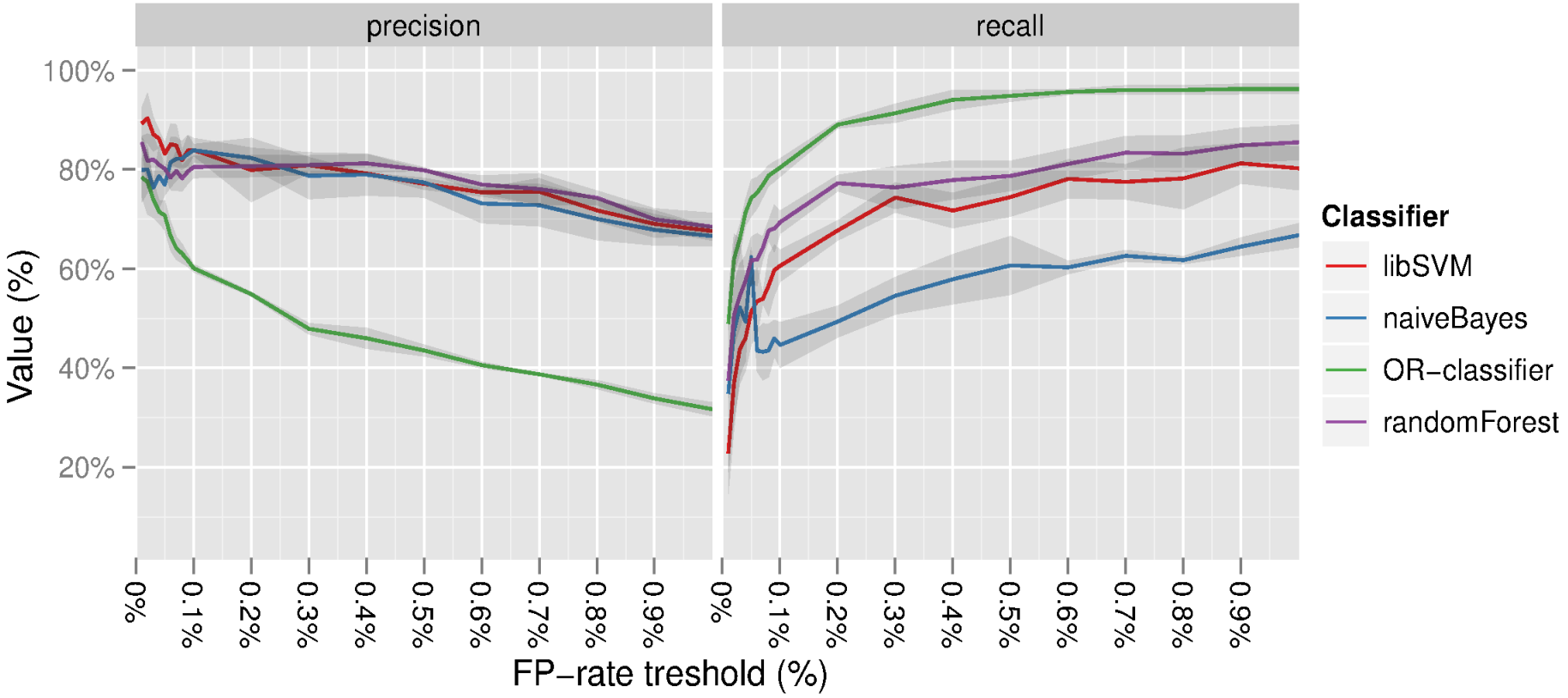


Masinõpe

- Erinevate mustrite väljundit on võimalik agregeerida.

	X ja **Y**	**X** ja **Y** kohtuvad	
Ilves ja **Obama** kohtuvad	1	1	õige
Ain ja **Jüri** einestasid	1	0	väär

Katse: NER isikunimi



1. aasta kokkuvõte

- Mustrite üldistamise algoritm ja tõestused.
- Implementatsioon (märgendaja, käsurea tööriistad)
- Katse isikunimede, organisatsioonide ja asukohtade tuvastamiseks.
- Aktiivõppe metoodika ja katse Wikipedia isikute ja nende sünniaegade tuvastamiseks; Twitteri postitustest sündmuste ja asukohtade tuvastamine.

Praegune ja tulevane töö

- Lahenduse kirjutamine C++ keeles parema efektiivsuse saamiseks.
- Masinõppe mudeli täiustamine.
  - Statistiliselt olulised mustrid.
  - Korpuste eeltöötlus ja lausete klasterdamine.
  - Leitud vastete järeltöötlus ning valepositiivsete tulemuste vähendamine.

Reaalsete rakenduste implementeerimine

- Olemasoleva A. Tkachenko NER lahenduse ühildamine meie meetodiga täpsuse parandamiseks.
- Tööriist lingvistidele statistiliselt huvitavate mustrite leidmiseks ja kuvamiseks märgendatud tekstidest.
- ...

Küsimused ja vastused