KV-järjestelmät suomelle

Lili Aunimo

lili.aunimo@cs.helsinki.fi

Helsingin yliopisto – Tietojenkäsittelytieteen laitos

Johdanto

- Lili Aunimo, Juha Makkonen ja Reeta Kuuskoski: Cross-Language Question Answering for Finnish
- Mikä on kysymysvastausjärjestelmä?
- Monikielinen KV-järjestelmä?
- Suppean alan KV-järjestelmä vai yleinen KV-järjestelmä?

Tässä esityksessä

- Ongelman esittely
- Ratkaisu:
  - Yksinkertainen järjestelmä: *Tikka*
  - Kieli- ja tietämysteknologiaa hyödyntävä *Varis*
  - Suomen kielen käsittely
  - Kysymyksen käsittely, tiedonhaku ja vastauksen käsittely
- Ratkaisun arviointi
- Johtopäätökset

Ongelma, Osa I

- Syöte:
  - Suomenkielinen kysymys
  - Milloin koelentäjä Chuck Yeager rikkoi äänivallin?
  - Kysymyksen kääntäminen englanniksi ja kysymyksen muuntaminen semanttiseen esitysmuotoon
- Tietokanta:
  - Englanninkielinen tekstidokumenttikokoelma
  - The Glasgow Herald ja Los Angeles Times, 670 Mt
  - Voidaanko käyttää perinteisiä tiedonhakumenetelmiä?

Ongelma, Osa II

- Tuloste:
  - Englanninkielinen vastaus
  - 1 LA102394.0299 Oct.14, 1947
  - Luottamusarvo, NIL-vastaukset
  - Miten vastaus eristetään dokumentista?

Motivaatio

- Tiedonhaku vs. dokumenttienhaku
- Monikielisyys
- Käyttöliittymärajoitukset, esim. pienet päätelaitteet ja puhesynteesi
- Suomen kieli

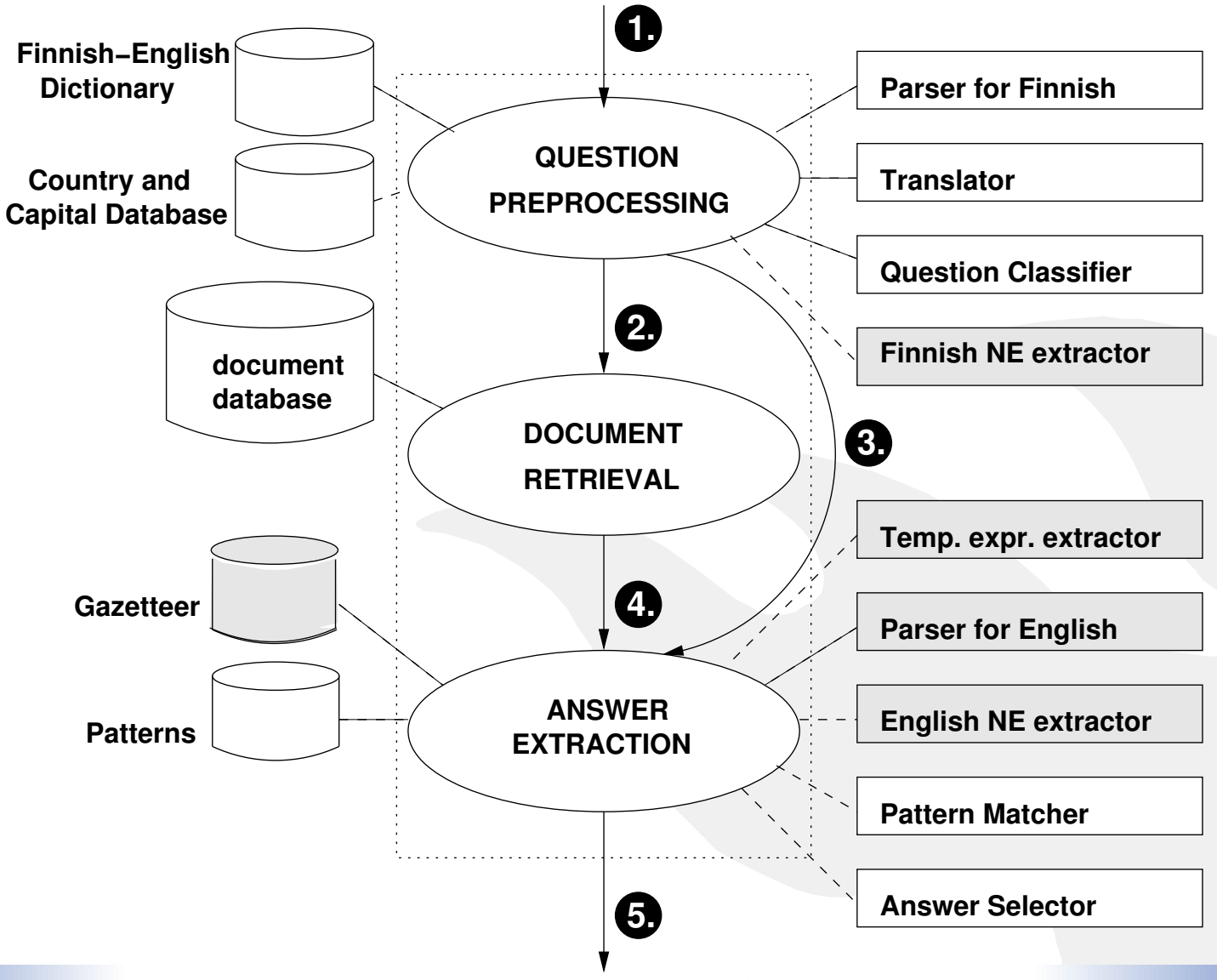
Suomen kieli, osa I

- Ensimmäinen yleinen KV-järjestelmä suomelle
- Morfologia
  - Kieliopillisten sijojen ja päätteiden runsaus, vokaaliharmonia ja astevaihtelu
  - Cannesistammekaan, salissa, säälistä, vesi, vedessä
  - Yhdyssanat
  - tuliaselaki: tuliase laki, kulttuuripääkaupunki: kulttuuri pääkaupunki
- Sanasto, osa I
  - immigrate tulla siirtolaisena
  - immigrer (fr), immigrare (it), imigrar (pt)

Suomen kieli, osa II

- Sanasto, osa II
  - Moniselitteisyys
  - Kerros: hiekkakerros, asuinkerros
  - Käännösmoniselitteisyys
  - Kerros: layer (kerros, kerrostuma, peite, taivukas), floor (lattia, istuntosali, kerros, tanssilattia, pohja)

Järjestelmäarkkitehtuuri



Kysymyksen käsittely

- Luokittelu
  - *henkilö*, *mitta*, *muu*, *organisatio*, *paikka*, *päivämäärä*
- Varis: Muunnos semanttiseen esitysmuotoon
  - *teema*, *fokus*
  - *teema*: koelentäjä Chuck Yeager, *fokus*: *päivämäärä*: äänivallin rikkoi
- Kääntäminen
  - *Tikka*: substantiivit ja adjektiiviattribuutit (*test pilot*, *sound barrier*, *sonic barrier*)
  - *Varis*: Kaikki sanat (*testi 3*), (*pilotti 1*), (*rikkoa 20*), (*äänivalli 2*). Yhteensä 120 kombinaatiota

Dokumenttienhaku, *Tikka*

- Hakukoneen kyselytermit: *test pilot Chuck Yeager sound barrier sonic barrier*
- Boolean-haun tulos: 1 dokumentti

For many, seeing Chuck Yeager – who made his historic supersonic flight Oct. 14, 1947 – was the highlight of this year's show, in which the Air Force gives taxpayers, aviation workers, curious motor heads and technological eggheads a ...

Dokumenttienhaku, *Varis*

- Haetaan erikseen kaikilla kombinaatioilla boolean-haulla, *120 kyselyä*.
- Tulos: 13 kyselyä tuotti vastauksia.
- Käännösten lukumäärä: *testi 2*, *pilotti 1*, *rikkoa 6*, *äänivalli 2*.
- Kontekstiheuristiikka liukuu vastausten läpi kolmen virkkeen kokoisessa ikkunassa ja valitsee ikkunat jatkokäsittelyyn.

Vastausten käsittely, *Varis*

- Vastausikkunakandidaattien esikäsittely

For many, seeing <ind>**Chuck Yeager**</ind> – who made his historic supersonic flight <temp s='19471014' e='19471014'>**Oct. 14, 1947**</temp> – was the highlight of <temp s='19940101' e='19941231'>**this year**</temp>'s show, in which <org>the Air Force</org> gives taxpayers, aviation workers, curious motor heads and technological eggheads a ...

Vastausten käsittely

- Vastaushahmojen sovittaminen tekstiin

```
Chuck Yeager [^\.\?\!]+
((Jan|Feb|Mar|Apr|Aug|Sep|Oct|Nov|Dec)\.
[1-9]{1,2}, [1-9]{4})
```

- Luottamusarvon laskeminen
- *Tikka*: Vastauksen frekvenssi ja suhteellinen osuus.
- *Varis*: Normalisoitu keskiarvo kysymyssanojen etäisyydestä konteksti-ikkunan lopusta ja vastauksen frekvenssi.

Arviointi

		Tikka		*Varis*	
type	**questions**	**patterns**	**%**	**patterns**	**%**
date	31	3	19.4	15	32.2
location	41	18	46.3	32	31.7
measure	34	22	32.4	22	38.2
person	39	16	17.9	11	23.1
object	5	0	0.0	0	0.0
organization	28	0	3.6	16	25.0
other	22	5	4.5	19	27.2
total	200	64	22.5	115	29.0

Arviointi

	Tikka		Varis	
	Absolute	**Percentage**	**Absolute**	**%**
Accuracy	45/200	22.5	58/200	29.0
NIL	132/200	66.0	75/200	37.5
Accuracy of NIL	13/132	9.8	12/75	16.0

Johtopäätökset

- *Tikka* ja *Varis* ovat ensimmäisiä KV-järjestelmiä suomen kielelle.
- Molemmat perustuvat:
  - kysymyksen luokitteluun,
  - kysymystermien kääntämiseen,
  - hakukoneen käyttöön ja
  - hahmonsovitukseen.
- Lisäksi *Varis* hyödyntää:
  - kysymyksen semanttista analyysiä,
  - käännösehdokkaiden karsintaa ja
  - vastausdokumenttien semanttista annotointia.