Sammendrag

INF5820 – H2008

Jan Tore Lønning

Institutt for Informatikk
Universitetet i Oslo

1. oktober

Outline

Outline

Hva?

- En forkortet versjon
- Få med det viktigste
- Tilpasset:
  - Brukere
  - Formål

Varianter

Ulike formål

- Sammendrag av ett eller flere dokument
- Et generelt sammendrag eller som svar på spørsmål/søk (“query-focused”)

Ulike typer

- Utdrag (“extract”)
- “Abstract” (bruker andre ord)

Skal vi gjøre det automatisk, blir det helst varianter av utdrag.

Trinnene

Fra generering av naturlige språk (NLG):

1. Velg innhold (hvilke setninger vil vi bruke)
2. Rekkefølge
3. Setningsrealisering

Outline

Viktigste termer

- Hvilke setninger er mest sentrale i en artikkel?
- Se på termer.
- Hvilke termer er viktige i et dokument?
- De som forekommer hyppig i dokumentet
- i forhold til hvor hyppig de ellers forekommer.

Fra IR (for to uker siden)

- Skal alle termer telle like mye?
- Idé: Termer som forekommer i få dokumenter er karakteristiske for de dokumentene hvor de forekommer
- N er antall dokumenter og n_i er de hvor term i forekommer
- IDF, **inverse document frequency**: $\frac{N}{n_i}$
- $$\text{idf}_i = \log\left(\frac{N}{n_i}\right)$$
- **td-idf**-vekting: $w_{i,j} = \text{tf}_{i,j} \times \text{idf}_i$
- der $\text{tf}_{i,j}$ antall forekomster av term i i dokumemt j
- Sentralt i flere anvendelser, som summering.

Annen vekting

- $$\text{weight}(w_i) = \begin{cases} 1 & \text{if } -2\log(\lambda(w_i)) > 10 \\ 0 & \text{ellers} \end{cases}$$
- $$\text{weight}(s_i) = \sum_{w \in s_i} \frac{\text{weight}(w)}{|\{w \mid w \in s_i\}|}$$
- Denne metoden kalles centroidesentrert (tyngdepunktsentrert).

Alternativ (“unsupervised”)

- Bruke cosinus-mål for å finne setninger som er sentrale i forhold til andre setninger i dokumentet.
- I stedet for termer bruk retorisk sturktur.
- Legge vekt på plassering i dokumentet: “Første setning i andre avsnitt”.

Treningsmateriale

Par av tekster og utdrag av tekstene

Mulige trekk

- Plassering i dokumentet.
- “Cue phrases”: *in summary*, *in this paper*...
- Viktige termer, som i det uovervåkete tilfellet.
- Setningslengde mm.

Trening

- Se på de setningene i artikkelen som også finnes i sammendraget.
- Hvilke trekk har de?

Dekoding

Gi setningene i en artikkel karakter ut i fra disse trekkene. Trekk ut de viktigste.

- Svakhet: trenger treningsmateriale som er utdrag. Setningenes må finnes i artikkelen eller vi må forbinde setninger i sammendraget med setninger i artikkelen før trening.

Outline

For ett dokument

- For ett dokument bruker vi setningene nesten uforandret.
- Men vil forkorte, fjerne uvesentligheter.
- Kanskje se på NP-ledd, pronomen, osv. for å få koherens.

Outline

- Hva blir annerledes?
- Samme informasjon i flere dokument

Valg av innhold

- Bare legge til ny informasjon.
- Alternativ 1(MMR):
  - En og en setning (som for enkeltdokument)
  - Men straffe setninger som er like de vi alt har valgt
- Alternativ 2 (“clustering”)
  - Trekke ut for mange kandidater.
  - Organisere dem i klynger med beslektede setninger.
  - Trekke representative medlemmer fra hver klynge.

Rekkefølge

Kronologi

- Med ett dokument, kan vi følge rekkefølgen i dokumentet, men ikke for flere dokumenter.
- En mulighet er å prøve å finne en kronologisk rekkefølge.
- Bra for narrative (etterretningsrapporter)
- Men ikke for alt.
- Kan være vanskelig.

Koherens

- Et minimumskrav
- Sammenheng med realisering

Setningsrealisering

- Unngå uklare overganger f.eks. *på den andre side* hvis setningen som sto foran ikke er med.
- Se spesielt på navn og pronomen.
- Unngå pronomen hvis referenten er uklar.
- Men ikke gjenta lange navn; blir oppstykket.
- Først navn, så pronomen.

Outline

- Et spesifikt informasjonsbehov, der vi ønsker et svar som
  - Er fyldigere enn i factoid-spørsmål, ikke bare et navn eller årstall.
  - Vi ønsker et svar, ikke et dokument som i IR
  - Vi har et spesifikt behov — ikke bare et generelt sammendrag
- Løsningere vil kombinere ulike metoder, f.eks.
  - Kombinere IR-metoder med metoder fra sammendrag for å finne innhold. Deretter som i sammendrag.
  - Kan også bruke metoder for informasjonsuttrekning (IE) (Kap.22)

Evaluering

- Liknende metoder (ROUGE) som for MT (BLEU)
- Vi skal se på BLEU mm. senere i semesteret.