SMT—Statistisk maskinoversettelse 2

INF5820 – H2008

Jan Tore Lønning

Institutt for Informatikk
Universitetet i Oslo

8. oktober

Outline

Outline

"Noisy channel" MT

- Skal finne den beste engelske oversettelsen $\hat{E}$ av en fremmed setning F.

$$\begin{aligned}\hat{E} &= \arg\max_E P(E \mid F) \\ &= \arg\max_E P(F \mid E)P(E)\end{aligned}$$

- Vi snur på flisa: betrakter F som en oversettelse (forvanskning) av E og spør hvilken E.

Modell

$$\hat{E} = \arg\max_{E} P(F \mid E)P(E)$$

- For språkmodellen $P(E) = P(e_1 e_2 \ldots e_n)$ kan vi bruke n-gram.
- Flere alternative modeller for $P(F|E)$:
  - For ordbasert oversettelse deler problemet seg i to:
    a) Ord-for-ord-delen. Sannsynligheten for at et ord er oversettelse av et annet.
    b) Plassering, “alignment” mellom et ord og dets oversettelse
  - Et annet alternativ er frasebaserte modeller.

IBM-modellene

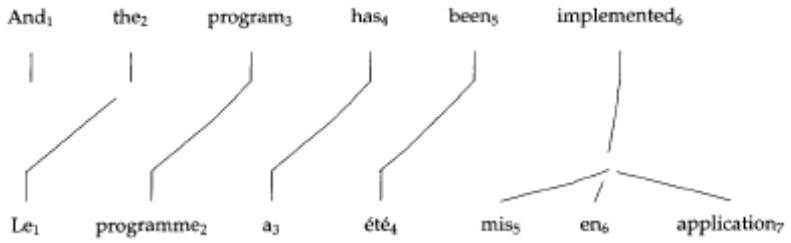
"Alignment"—felles for modellene

- Hvert ord i F stammer fra nøyaktig ett ord i E, eller fra ingen.
- Ett ord i E går til 0 eller flere ord i F.
- (Ett ord i F kan ikke svare til flere ord i E.)

Modell 1

- Alle alignments er like sannsynlige.
- I stedet for å måtte se på alle $((k+1)^m)$ mulige alignments,
- kan vi se på hvert ord i F, og hvilket ord i E det kommer fra, uavhengig av de andre ordene i F
- Vi betrakter $t(f_j \mid E) = \sum_{i=0}^{k} t(f_j \mid e_i)$

Representasjon



Denne alignmenten:

<2, 3, 4, 5, 6, 6, 6>

<a1,a2,a3,a4,a5,a6>

Generelt:

- Lengden av engelsk streng: k
- Lengden av fransk streng: m
- En alignment er en vektor av m tall, hvert mellom 0 og k. (Hvorfor 0?)
- (k+1)^m mange forskjellige

OBS:

- I prinsippet kan flere franske ord stamme fra samme engelske
- Men hvert fransk ord stammer enten fra et engelsk eller fra ingenting

IBM-modellene

IBM-modell 1 og 2

$$\Pr(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$$

$$\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \Pr(m \mid \mathbf{e}) \prod_{j=1}^{m} \Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) \Pr(f_j \mid a_1^{j}, f_1^{j-1}, m, \mathbf{e})$$

Betyr:

- Velg lengde av den franske strengen gitt den engelske.
- På plass 1 i den franske strengen, velg hvilken plass i den engelske vi skal se på gitt den engelske og lengden av den franske.
- Velg deretter det franske ordet på grunnlag av denne etablerte forbindelsen og de samme opplysningene

…..

- Se på neste plass i den franske strengen. Velg plass i den engelske strengen på grunnlag av den engelske, lengden av den franske og alle fobindleser og ord som er valgt så langt.
- Ta også dette med i betraktning og velg hvilket fransk ord som skal stå her.

Så langt ikke en tilnærming.

IBM-modell 1

$$\Pr(\mathbf{f},\mathbf{a}\mid\mathbf{e}) = \Pr(m\mid\mathbf{e})\prod_{j=1}^{m}\Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, \mathbf{e})\Pr(f_j \mid a_1^{j}, f_1^{j-1}, m, \mathbf{e})$$

Nå skal vi gjøre approksimasjoner.

$\Pr(m\mid\mathbf{e})$ er uavhengig av m og **e**.

$\Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) = (k+1)^{-1}$, dvs den avhenger bare av lengden k av **e**.

$\Pr(f_j \mid a_1^{j}, f_1^{j-1}, m, \mathbf{e}) = t(f_j \mid e_{a_j})$, oversettelsessannsynligheten av f_j, gitt e_{a_j}, dvs den avhenger bare av det ordet den er forbundet med, jfr tagging.

Hadde vi hatt et word-aligned korpus kunne vi estimert denne ved opptelling:

$$t(f_j \mid e_{a_j}) = \frac{C(f_j, e_{a_j})}{\sum_f C(f, e_{a_j})}$$

$$\Pr(\mathbf{f},\mathbf{a}\mid\mathbf{e}) = \varepsilon\prod_{j=1}^{m}(k+1)^{-1}t(f_j \mid e_{a_j})$$

$$\Pr(\mathbf{f},\mathbf{a}\mid\mathbf{e}) = \frac{\varepsilon}{(k+1)^m}\prod_{j=1}^{m}t(f_j \mid e_{a_j})$$

Vi har

$$\Pr(\mathbf{f},\mathbf{a}\mid\mathbf{e}) = \frac{\varepsilon}{(k+1)^m}\prod_{j=1}^{m} t(f_j \mid e_{a_j})$$

Da vil

$$\Pr(\mathbf{f}\mid\mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f},\mathbf{a}\mid\mathbf{e}) = \sum_{\mathbf{a}} \frac{\varepsilon}{(k+1)^m}\prod_{j=1}^{m} t(f_j \mid e_{a_j})$$

$$\Pr(\mathbf{f}\mid\mathbf{e}) = \sum_{a_1=0}^{k}\cdots\sum_{am=0}^{k} \frac{\varepsilon}{(k+1)^m}\prod_{j=1}^{m} t(f_j \mid e_{a_j})$$

$$\Pr(\mathbf{f}\mid\mathbf{e}) = \frac{\varepsilon}{(k+1)^m}\prod_{j=1}^{m}\sum_{i=0}^{k} t(f_j \mid e_i)$$

Outline

Hvis bare . . .

- Gitt et aligned korpus. Da kan vi finne oversettelsessannsynligheten for ord ved ren opptelling.
- Omvendt: gitt oversettelsessannsynligheter kan vi finne alignment-sannsynligheter:

$$P(A, F \mid E) = P(A \mid F, E) \times P(F|E) \text{(produktregelen)}$$

$$P(A \mid F, E) = \frac{P(A, F \mid E)}{P(F|E)}$$

$$\text{der } P(F \mid E) = \sum_A P(A, F \mid E)$$

EM-algoritmen

Simultan læring av A og ordoversettelsessannsynligheter.

1. Start med uniform fordeling av $t(f_j \mid e_i)$.
2. Bruk t til å beregne $P(A, F \mid E)$ for hvert par av "sentence aligned" setninger $E - F$ og enhver mulig "word alignment" av dem.
3. Normaliser $P(A, F \mid E)$ for å få $P(A \mid E, F)$
4. Tell opp "fractional counts": $tc(f_j \mid e_i) = \sum_{\{A \mid A(f_j)=e_i\}} P(A \mid E, F)$
5. Normaliser med $tc(\cdot \mid e_i)$ for å få ny $t(f_j \mid e_i)$
6. Gjenta trinn 2–5 til maskinen går varm

Outline

Husk

$$\hat{E} = \arg\max_{E} P(F \mid E)P(E)$$

- For språkmodellen $P(E) = P(e_1 e_2 \ldots e_n)$ kan vi bruke n-gram.
- Flere alternative modeller for $P(F|E)$:
  - For ordbasert oversettelse deler problemet seg i to:
    a) Ord-for-ord-delen. Sannsynligheten for at et ord er oversettelse av et annet.
    b) Plassering, "alignment" mellom et ord og dets oversettelse
  - Et annet alternativ er frasebaserte modeller.
- HMM-alignment er et alternativ til IBM-modellene på dette punktet.

ORD-aligned-oversettelse

$$\Pr(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$$

$$\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \Pr(m \mid \mathbf{e}) \prod_{j=1}^{m} \Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) \Pr(f_j \mid a_1^{j}, f_1^{j-1}, m, \mathbf{e})$$

Dette er: **eksakt** – ikke en tilnærming
Det er felles for IBM og HMM-alignment

Betyr:

- Velg lengde av den franske strengen gitt den engelske.
- På plass 1 i den franske strengen, velg hvilken plass i den engelske vi skal se på gitt den engelske og lengden av den franske.
- Velg deretter det franske ordet på grunnlag av denne etablerte forbindelsen og de samme opplysningene

…..

- Se på neste plass i den franske strengen. Velg plass i den engelske strengen på grunnlag av den engelske, lengden av den franske og alle fobindleser og ord som er valgt så langt.
- Ta også dette med i betraktning og velg hvilket fransk ord som skal stå her.

IBM-modell 1

$$\Pr(\mathbf{f},\mathbf{a}\mid\mathbf{e}) = \Pr(m\mid\mathbf{e})\prod_{j=1}^{m}\Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, \mathbf{e})\Pr(f_j \mid a_1^{j}, f_1^{j-1}, m, \mathbf{e})$$

Approksimasjoner.

1. $\Pr(m\mid\mathbf{e})$ er uavhengig av m og **e**.
2. $\Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) = (k+1)^{-1}$, dvs den avhenger bare av lengden k av **e**.
3. $\Pr(f_j \mid a_1^{j}, f_1^{j-1}, m, \mathbf{e}) = t(f_j \mid e_{a_j})$, oversettelsessannsynligheten av f_j, gitt e_{a_j}, dvs den avhenger bare av det ordet den er forbundet med, jfr tagging.

HMM-alignment

pkt. 3 som IBM

pkt.1: P(J | I) der J er lengden av F og I lengden av E.

pkt.2: $\Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) = \Pr(a_j \mid a_{j-1}, I)$

mao:
a_j avhenger bare av de andre a_k –ene, ikke av f_k –ene

a_j avhenger bare av a_(j-1) og av lengden I av E

Markov-betingelser

Hopp-sannsynlighet

- $P(a_j \mid a_{j-1}, l)$
- Vi er interessert i hvor sannsynlig det er at a_j flytter hvor langt fra a_{j-1}
- Dette kan avhenge av l (flytter lengre i en lang setning),
- Men lite avhengig av a_{j-1} hvor langt det flyttes.
- Derfor vil vi se på $P(|a_i - a_{i-1}| \mid l)$

$$\frac{c(i - i')}{\sum_{i''=1}^{l} c(i'' - i')}$$

Outline

Begrensninger i de ordbaserte-modellene:

- Ett ord kan ikke komme fra flere ord: *the car* $\rightarrow$*bilen*.
- Og selv om ett ord kan gå til flere gjøres det ikke nyanser.
  - Hvis *bilen* $\rightarrow$ *the car*, tolkes det som
  - *bilen* $\rightarrow$ *the*, og
  - *bilen* $\rightarrow$ *car*
- I mange tilfeller er det naturlig mange-til-mange korrespondanse.

Frasebasert modell

$$P(F \mid E) = \prod_{i=1}^{I} \phi(\overline{f}_i, \overline{e}_i) d(a_i - b_{i-1})$$

- Oversettelsessannsynlighet: $(\overline{f}_i, \overline{e}_i)$
- "distortion" sannsynlighet $d(a_i - b_{i-1})$, der b_{i-1} er slutten på forrige frase og a_i begynnelsen på denne.
- Kommentarer:
  - Kurant modell for dekoding.
  - Vanskelig å bruke direkte i parameterlæring: "For mange mulige alignements"

Frasebasert parameterlæring

- Ett alternativ:
  - Lag ord-alignment fra E til F
  - og fra F til E
  - Ta det som er felles fra de to
  - og prøv å bygg ut med deler fra de to til en full frase-alignement.
  - Betrakt også alle mulige underfraser av en slik som mulige kandidater.

Outline

Dekoding

$$\hat{E} = \arg\max_{E} P(F \mid E) P(E)$$

- Vi må se på de to delene simultant.
- Vi kan ikke prøve alle mulige oversettelser E.
- Første avskjæring: En tabell av fraser (i språket E) som svarer til fraser i setningen F med en rimelig sannsynlighet.

Søk

- Søkerom:
  - Mulige kombinasjoner av slike fraser
  - Merk av hvilke deler av F som er oversatt
  - Beregn sannsynligheter for de oversettelsene som vleges
- Også dette er for stort: Avskjær underveis på grunnlag av
  - Reell kostnad av hva som er gjort så langt: oversettelsessannynlighet og språkmodell
  - Estimert kostnad av gjenstående: oversettelsessannsynlighet