SyMGiza++: A Tool for Parallel Computation of Symmetrized Word Alignment Models

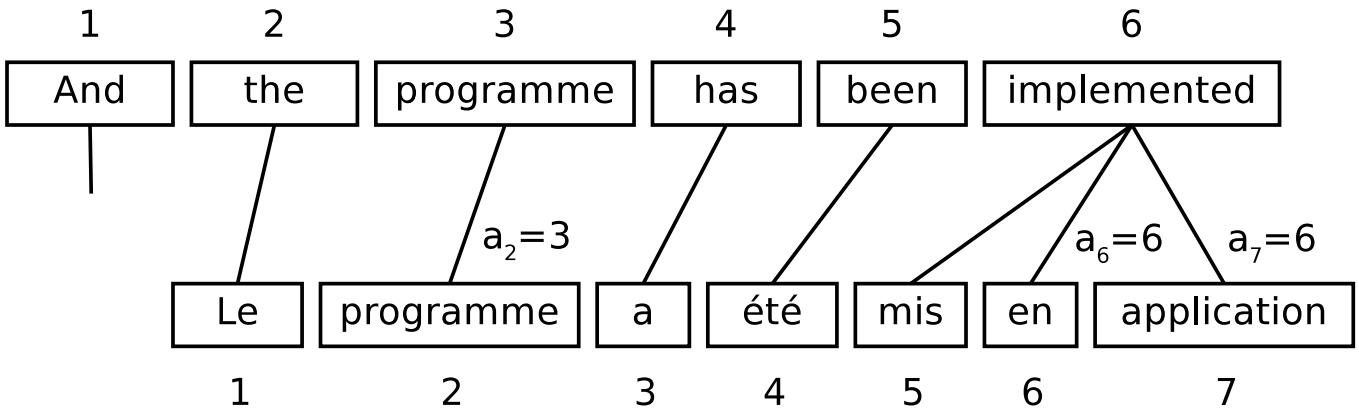
Marcin Junczys-Dowmunt and Arkadiusz Szał

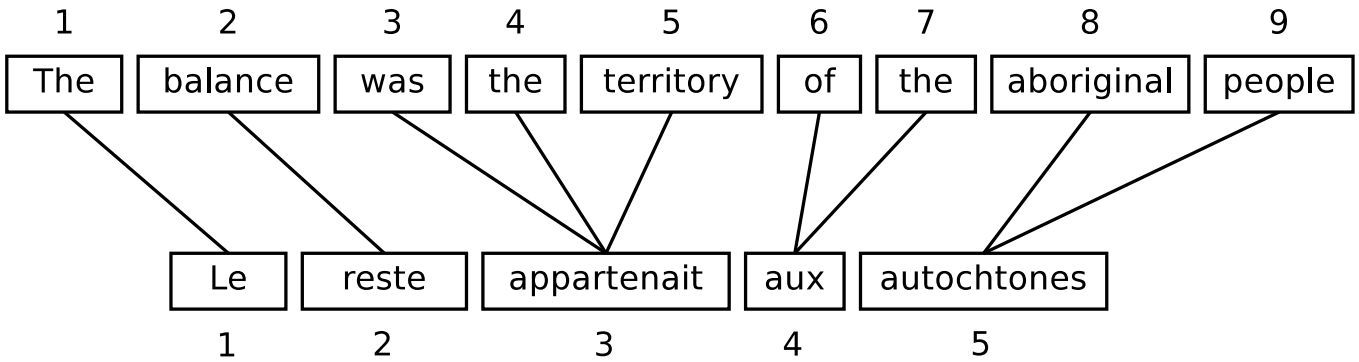
Information Systems Laboratory, Adam Mickiewicz University, Poznań, Poland

Oct. 19, 2010

Introduction

Directed word alignment models





Giza++

- **Giza++** implements maximum likelihood estimators for several statistical alignment models (IBM Model 1 through 5, HMM alignment model, Model 6).
- The EM algorithm is used for the estimation of model parameters. It consists of two steps iterativly repeated steps:
  - E-step — a previously computed model is applied to the data. The expected counts for specific parameters are collected.
  - M-step — the expected counts are taken as fact and used to estimate the probabilities of the next model.

Giza++

- **Giza++** implements maximum likelihood estimators for several statistical alignment models (IBM Model 1 through 5, HMM alignment model, Model 6).
- The EM algorithm is used for the estimation of model parameters. It consists of two steps iterativly repeated steps:
  - E-step — a previously computed model is applied to the data. The expected counts for specific parameters are collected.
  - M-step — the expected counts are taken as fact and used to estimate the probabilities of the next model.

Example modification of IBM Model 1 (I)

Two **directed** alignment models — Pr_α and Pr_β — are trained **in parallel**.

$$Pr_\alpha(\mathbf{f}|\mathbf{e}) = \frac{\epsilon(m|l)}{(l+1)^m} \sum_{\mathbf{a}} \prod_{j=1}^{m} t_\alpha(f_j|e_{a_j})$$

The **translation probabilities** t_α — free parameters of model Pr_α — are estimated as:

$$t_\alpha(f|e) = \frac{\sum_{s=1}^{S} c(f|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)})}{\sum_{f'}\sum_{s=1}^{S} c(f'|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)})},$$

$$c(f|e;\mathbf{f},\mathbf{e}) = \sum_{\mathbf{a}} Pr_\alpha(\mathbf{a}|\mathbf{f},\mathbf{e}) \sum_{i,j} \delta(f,f_j)\delta(e,e_i),$$

$$Pr_\alpha(\mathbf{a}|\mathbf{f},\mathbf{e}) = \frac{\prod_{j=1}^{m} \tilde{t}_\alpha(f_j|e_{a_j})}{\sum_{\mathbf{a}}\prod_{j=1}^{m} \tilde{t}_\alpha(f_j|e_{a_j})},$$

where $\tilde{t}_\alpha$ are translation probabilities from the **previous iteration**.

Example modification of IBM Model 1 (I)

Two **directed** alignment models — Pr_α and Pr_β — are trained **in parallel**.

$$Pr_\alpha(\mathbf{f}|\mathbf{e}) = \frac{\epsilon(m|l)}{(l+1)^m} \sum_{\mathbf{a}} \prod_{j=1}^{m} t_\alpha(f_j|e_{a_j})$$

The **translation probabilities** t_α — free parameters of model Pr_α — are estimated as:

$$t_\alpha(f|e) = \frac{\sum_{s=1}^{S} c(f|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)})}{\sum_{f'}\sum_{s=1}^{S} c(f'|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)})},$$

$$c(f|e;\mathbf{f},\mathbf{e}) = \sum_{\mathbf{a}} Pr_\alpha(\mathbf{a}|\mathbf{f},\mathbf{e}) \sum_{i,j} \delta(f,f_j)\delta(e,e_i),$$

$$Pr_\alpha(\mathbf{a}|\mathbf{f},\mathbf{e}) = \frac{\prod_{j=1}^{m} \tilde{t}_\alpha(f_j|e_{a_j})}{\sum_{\mathbf{a}}\prod_{j=1}^{m} \tilde{t}_\alpha(f_j|e_{a_j})},$$

where $\tilde{t}_\alpha$ are translation probabilities from the **previous iteration**.

Results

Alignment Method	Time [m]	Prec [%]	Rec [%]	AER [%]
Giza++ en-fr	–	91.19	92.20	8.39
Giza++ fr-en	–	91.82	87.96	9.79
Giza++ refined	457	93.24	92.59	7.02
SyMGiza++ en-fr	–	94.22	93.88	**5.92**
SyMGiza++ fr-en	–	95.27	88.58	7.57
SyMGiza++ refined	332	**94.34**	**94.08**	**5.76**

The last slide

Thank you very much!