Lexical alignment: IBM models 1 and 2

MLE via EM for categorical distributions

Wilker Aziz

April 10, 2017

Translation data

Let's assume we are confronted with a new language
and luckily we managed to obtain some sentence-aligned data

the black dog	$\square$ $\circledast$
the nice dog	$\square$ $\cup$
the black cat	$\boxdot$ $\circledast$
a dog chasing a cat	$\boxdot$ $\triangleleft$ $\square$

Translation data

Let's assume we are confronted with a new language
and luckily we managed to obtain some sentence-aligned data

the black dog	$\square \circledast$
the nice dog	$\square \cup$
the black cat	$\boxdot \circledast$
a dog chasing a cat	$\boxdot \triangleleft \square$

Is there anything we could say about this language?

Translation by analogy

the black dog	$\square\ \circledast$
the nice dog	$\square\ \cup$
the black cat	$\boxdot\ \circledast$
a dog chasing a cat	$\boxdot \triangleleft \square$

A few hypotheses:

Translation by analogy

the black dog	$\Box\ \circledast$
the nice dog	$\Box\ \cup$
the black cat	$\boxdot\ \circledast$
a dog chasing a cat	$\boxdot\ \triangleleft\ \Box$

A few hypotheses:

- $\Box \iff$ dog

Translation by analogy

the black dog	$\square\ \circledast$
the nice dog	$\square\ \cup$
the black cat	$\boxdot\ \circledast$
a dog chasing a cat	$\boxdot\ \triangleleft\ \square$

A few hypotheses:

- $\square \iff$ dog
- $\boxdot \iff$ cat

Translation by analogy

the black dog	$\square\ \circledast$
the nice dog	$\square\ \cup$
the black cat	$\boxdot\ \circledast$
a dog chasing a cat	$\boxdot \triangleleft \square$

A few hypotheses:

- $\square \iff$ dog
- $\boxdot \iff$ cat
- $\circledast \iff$ black

Translation by analogy

the black dog	$\square\ \circledast$
the nice dog	$\square\ \cup$
the black cat	$\boxdot\ \circledast$
a dog chasing a cat	$\boxdot\ \triangleleft\ \square$

A few hypotheses:

- $\square \iff$ dog
- $\boxdot \iff$ cat
- $\circledast \iff$ black
- nouns seem to preceed adjectives

Translation by analogy

the black dog	$\square \circledast$
the nice dog	$\square \cup$
the black cat	$\boxdot \circledast$
a dog chasing a cat	$\boxdot \triangleleft \square$

A few hypotheses:

- $\square \iff$ dog
- $\boxdot \iff$ cat
- $\circledast \iff$ black
- nouns seem to preceed adjectives
- determines are probably not expressed

Translation by analogy

the black dog	$\square\ \circledast$
the nice dog	$\square\ \cup$
the black cat	$\boxdot\ \circledast$
a dog chasing a cat	$\boxdot\ \triangleleft\ \square$

A few hypotheses:

- $\square \iff$ dog
- $\boxdot \iff$ cat
- $\circledast \iff$ black
- nouns seem to preceed adjectives
- determines are probably not expressed
- *chasing* may be expressed by $\triangleleft$
  and perhaps this language is OVS

Translation by analogy

the black dog	$\square\ \circledast$
the nice dog	$\square\ \cup$
the black cat	$\boxdot\ \circledast$
a dog chasing a cat	$\boxdot \triangleleft \square$

A few hypotheses:

- $\square \iff$ dog
- $\boxdot \iff$ cat
- $\circledast \iff$ black
- nouns seem to preceed adjectives
- determines are probably not expressed
- *chasing* may be expressed by $\triangleleft$
  and perhaps this language is OVS
- or perhaps *chasing* is realised by a verb with swapped arguments

Probabilistic lexical alignment models

This lecture is about operationalising this intuition

- through a probabilistic learning algorithm
- for a non-probabilistic approach see for example [Lardilleux and Lepage, 2009]

Content

Word-to-word alignments

Imagine you are given a text

the black dog	o cão preto
the nice dog	o cão amigo
the black cat	o gato preto
a dog chasing a cat	um cão perseguindo um gato

Word-to-word alignments

Now imagine the French words were replaced by placeholders

the black dog	$F_1\ F_2\ F_3$
the nice dog	$F_1\ F_2\ F_3$
the black cat	$F_1\ F_2\ F_3$
a dog chasing a cat	$F_1\ F_2\ F_3\ F_4\ F_5$

Word-to-word alignments

Now imagine the French words were replaced by placeholders

the black dog	$F_1\ F_2\ F_3$
the nice dog	$F_1\ F_2\ F_3$
the black cat	$F_1\ F_2\ F_3$
a dog chasing a cat	$F_1\ F_2\ F_3\ F_4\ F_5$

and suppose our task is to have a model explain the original data

Word-to-word alignments

Now imagine the French words were replaced by placeholders

the black dog	$F_1 \ F_2 \ F_3$
the nice dog	$F_1 \ F_2 \ F_3$
the black cat	$F_1 \ F_2 \ F_3$
a dog chasing a cat	$F_1 \ F_2 \ F_3 \ F_4 \ F_5$

and suppose our task is to have a model explain the original data
by generating each French word from exactly one English word

Generative story

For each sentence pair independently,

1. observe an English sentence $e_1, \cdots, e_m$
   and a French sentence length n
2. for each French word position j from 1 to n
   2.1 select an English position a_j
   2.2 conditioned on the English word e_{a_j}, generate f_j

Generative story

For each sentence pair independently,

1. observe an English sentence $e_1, \cdots, e_m$
   and a French sentence length n
2. for each French word position j from 1 to n
   2.1 select an English position a_j
   2.2 conditioned on the English word e_{a_j}, generate f_j

We have introduced an alignment
which is not directly visible in the data

Data augmentation

Observations:

the black dog | o cão preto

Imagine data is made of pairs: (a_j, f_j) and $e_{a_j} \to f_j$

Data augmentation

Observations:

the black dog | o cão preto

Imagine data is made of pairs: (a_j, f_j) and $e_{a_j} \to f_j$

the black dog $\Big|$ $(A_1, E_{A_1} \to F_1)$ $(A_2, E_{A_2} \to F_2)$ $(A_3, E_{A_3} \to F_3)$

Data augmentation

Observations:

$$\text{the black dog} \mid \text{o cão preto}$$

Imagine data is made of pairs: (a_j, f_j) and $e_{a_j} \to f_j$

$$\text{the black dog} \mid (1, E_{A_1} \to F_1)\ (A_2, E_{A_2} \to F_2)\ (A_3, E_{A_3} \to F_3)$$

Data augmentation

Observations:

$$\text{the black dog} \mid \text{o cão preto}$$

Imagine data is made of pairs: (a_j, f_j) and $e_{a_j} \to f_j$

$$\text{the black dog} \mid (1, \text{the} \to \text{o})\ (A_2, E_{A_2} \to F_2)\ (A_3, E_{A_3} \to F_3)$$

Data augmentation

Observations:

$$\text{the black dog} \mid \text{o cão preto}$$

Imagine data is made of pairs: (a_j, f_j) and $e_{a_j} \to f_j$

$$\text{the black dog} \mid (1, \text{the} \to \text{o})\ (3, E_{A_2} \to F_2)\ (A_3, E_{A_3} \to F_3)$$

Data augmentation

Observations:

the black dog | o cão preto

Imagine data is made of pairs: (a_j, f_j) and $e_{a_j} \to f_j$

the black dog | $(1, \text{the} \to \text{o})$ $(3, \text{dog} \to \text{cão})$ $(A_3, E_{A_3} \to F_3)$

Data augmentation

Observations:

the black dog | o cão preto

Imagine data is made of pairs: (a_j, f_j) and $e_{a_j} \to f_j$

the black dog | $(1, \text{the} \to \text{o})$ $(3, \text{dog} \to \text{cão})$ $(2, E_{A_3} \to F_3)$

Data augmentation

Observations:

the black dog | o cão preto

Imagine data is made of pairs: (a_j, f_j) and $e_{a_j} \to f_j$

the black dog | $(1, \text{the} \to \text{o})$ $(3, \text{dog} \to \text{cão})$ $(2, \text{black} \to \text{preto})$

Data augmentation

Observations:

$$\text{the black dog} \mid \text{o cão preto}$$

Imagine data is made of pairs: (a_j, f_j) and $e_{a_j} \to f_j$

$$\text{the black dog} \mid (1, \text{the} \to \text{o})\ (3, \text{dog} \to \text{cão})\ (2, \text{black} \to \text{preto})$$

$$\text{the black dog} \mid (1, \text{the} \to \text{o})\ (1, \text{the} \to \text{cão})\ (1, \text{the} \to \text{preto})$$

Data augmentation

Observations:

the black dog | o cão preto

Imagine data is made of pairs: (a_j, f_j) and $e_{a_j} \to f_j$

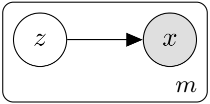
the black dog | $(1, \text{the} \to \text{o})$ $(3, \text{dog} \to \text{cão})$ $(2, \text{black} \to \text{preto})$

the black dog | $(1, \text{the} \to \text{o})$ $(1, \text{the} \to \text{cão})$ $(1, \text{the} \to \text{preto})$

the black dog | $(a_1, e_{a_1} \to f_1)$ $(a_2, e_{a_2} \to f_2)$ $(a_3, e_{a_3} \to f_3)$

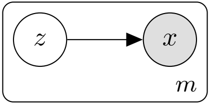
Content

Mixture models



- c mixture components
- each defines a distribution over the same data space $\mathcal{X}$
- plus a distribution over components themselves

Mixture models

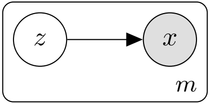


- c mixture components
- each defines a distribution over the same data space $\mathcal{X}$
- plus a distribution over components themselves

Generative story

1. select a mixture component $z \sim P(Z)$
2. generate an observation from it $x \sim P(X|Z = z)$

Mixture models



- c mixture components
- each defines a distribution over the same data space $\mathcal{X}$
- plus a distribution over components themselves

Marginal likelihood

$$P(x_1^m) = \prod_{i=1}^{m} \sum_{z=1}^{c} P(X = x_i, Z = z) \tag{1}$$

$$= \prod_{i=1}^{m} \sum_{z=1}^{c} P(Z = z) P(X = x_i | Z = z) \tag{2}$$

Interpretation

Missing data

- Let Z take one of c mixture components
- Assume data consists of pairs (x, z)
- x is always observed
- y is always missing

Interpretation

Missing data

- Let Z take one of c mixture components
- Assume data consists of pairs (x, z)
- x is always observed
- y is always missing

Inference: posterior distribution over possible Z for each x

$$P(Z = z|X = x) = \frac{P(Z = z, X = x)}{\sum_{z'=1}^{c} P(Z = z', X = x)} \tag{3}$$

$$= \frac{P(Z = z)P(X = x|Z = z)}{\sum_{z'=1}^{c} P(Z = z')P(X = x|Z = z')} \tag{4}$$

Non-identifiability

Different parameter settings, same distribution

Suppose $\mathcal{X} = \{a, b\}$ and $c = 2$
and let $P(Z = 1) = P(Z = 2) = 0.5$

Z	$X = a$	$X = b$
1	0.2	0.8
2	0.7	0.3
$P(X)$	0.45	0.55

Z	$X = a$	$X = b$
1	0.7	0.3
2	0.2	0.8
$P(X)$	0.45	0.55

Non-identifiability

Different parameter settings, same distribution

Suppose $\mathcal{X} = \{a, b\}$ and $c = 2$
and let $P(Z = 1) = P(Z = 2) = 0.5$

Z	$X = a$	$X = b$
1	0.2	0.8
2	0.7	0.3
$P(X)$	0.45	0.55

Z	$X = a$	$X = b$
1	0.7	0.3
2	0.2	0.8
$P(X)$	0.45	0.55

Problem for parameter estimation by hillclimbing

Maximum likelihood estimation

Suppose a dataset $\mathcal{D} = \{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$

Maximum likelihood estimation

Suppose a dataset $\mathcal{D} = \{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$
Suppose $P(X)$ is one of a parametric family with parameters θ

Maximum likelihood estimation

Suppose a dataset $\mathcal{D} = \{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$
Suppose $P(X)$ is one of a parametric family with parameters θ
Likelihood of iid observations

$$P(\mathcal{D}) = \prod_{i=1}^{m} P_\theta(X = x^{(i)})$$

Maximum likelihood estimation

Suppose a dataset $\mathcal{D} = \{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$
Suppose $P(X)$ is one of a parametric family with parameters θ
Likelihood of iid observations

$$P(\mathcal{D}) = \prod_{i=1}^{m} P_\theta(X = x^{(i)})$$

the score function is

$$l(\theta) = \sum_{i=1}^{m} \log P_\theta(X = x^{(i)})$$

Maximum likelihood estimation

Suppose a dataset $\mathcal{D} = \{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$
Suppose $P(X)$ is one of a parametric family with parameters θ
Likelihood of iid observations

$$P(\mathcal{D}) = \prod_{i=1}^{m} P_\theta(X = x^{(i)})$$

the score function is

$$l(\theta) = \sum_{i=1}^{m} \log P_\theta(X = x^{(i)})$$

then we choose

$$\theta^\star = \arg\max_\theta l(\theta)$$

MLE for categorical: estimation from fully observed data

Suppose we have **complete data**

- $\mathcal{D}_{\text{complete}} = \{(x^{(1)}, z^{(1)}), \ldots, (x^{(m)}, z^{(m)})\}$

MLE for categorical: estimation from fully observed data

Suppose we have **complete data**

- $\mathcal{D}_{\text{complete}} = \{(x^{(1)}, z^{(1)}), \ldots, (x^{(m)}, z^{(m)})\}$

Then, for a **categorical distribution**

$$P(X = x | Z = z) = \theta_{z,x}$$

and $n(z, x | \mathcal{D}_{\text{complete}}) =$ *count of* (z, x) *in* $\mathcal{D}_{\text{complete}}$

MLE solution:

$$P(X = x | Z = z) = \theta_{z,x} = \frac{n(z, x | \mathcal{D}_{\text{complete}})}{\sum_{x'} n(z, x' | \mathcal{D}_{\text{complete}})}$$

MLE for categorical: estimation from incomplete data

Expectation-Maximisation algorithm [Dempster et al., 1977]

E-step:

- for every observation x, imagine that every possible latent assignment z happened with probability $P_\theta(Z = z|X = x)$

$$\mathcal{D}_{\text{completed}} = \{(x, Z = 1), \ldots, (x, Z = c) : x \in \mathcal{D}\}$$

MLE for categorical: estimation from incomplete data

Expectation-Maximisation algorithm [Dempster et al., 1977]

M-step:

- reestimate θ as to climb the likelihood surface
- for categorical distributions $P(X = x|Z = z) = \theta_{z,x}$
  z and x are categorical
  $0 \leq \theta_{z,x} \leq 1$ and $\sum_{x \in X} \theta_{z,x} = 1$

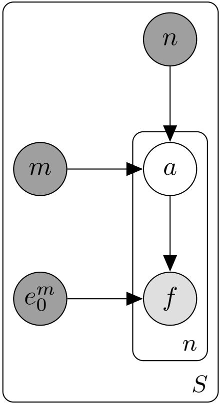
$$\theta_{z,x} = \frac{\mathbb{E}[n(z \to x|\mathcal{D}_{\text{completed}})]}{\sum_{x'} \mathbb{E}[n(z \to x'|\mathcal{D}_{\text{completed}})]} \tag{5}$$

$$= \frac{\sum_{i=1}^{m} \sum_{z'} P(z'|x^{(i)}) \mathbb{1}_z(z') \mathbb{1}_x(x^{(i)})}{\sum_{i=1}^{m} \sum_{x'} \sum_{z'} P(z'|x^{(i)}) \mathbb{1}_z(z') \mathbb{1}_{x'}(x^{(i)})} \tag{6}$$

$$= \frac{\sum_{i=1}^{m} P(z|x^{(i)}) \mathbb{1}_x(x^{(i)})}{\sum_{i=1}^{m} \sum_{x'} P(z|x^{(i)}) \mathbb{1}_{x'}(x^{(i)})} \tag{7}$$

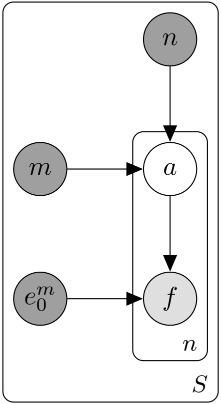
Content

IBM1: a constrained mixture model



Constrained mixture model

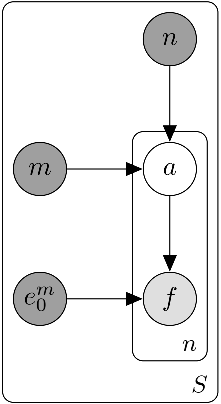
IBM1: a constrained mixture model



Constrained mixture model

- mixture components are English words

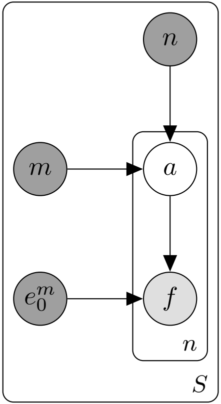
IBM1: a constrained mixture model



Constrained mixture model

- mixture components are English words
- but only English words that appear in the English sentence can be assigned

IBM1: a constrained mixture model



Constrained mixture model

- mixture components are English words
- but only English words that appear in the English sentence can be assigned
- a_j acts as an indicator for the mixture component that generates French word f_j
- e_0 is occupied by a special $\textsc{Null}$ component

Parameterisation

Alignment distribution: uniform

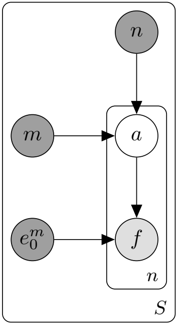
$$P(A|M = m, N = n) = \frac{1}{m + 1} \tag{8}$$

Lexical distribution: categorical

$$P(F|E = e) = \mathrm{Cat}(F|\theta_e) \tag{9}$$

- where $\theta_e \in \mathbb{R}^{v_F}$
- $0 \leq \theta_{e,f} \leq 1$
- $\sum_f \theta_{e,f} = 1$

IBM1: marginal likelihood



Marginal likelihood

$$P(f_1^n|e_0^m) = \sum_{a_1=0}^{m} \cdots \sum_{a_n=0}^{m} P(f_1^n, a_1^n|e_{a_j}) \tag{10}$$

$$= \sum_{a_1=0}^{m} \cdots \sum_{a_n=0}^{m} \prod_{j=1}^{n} P(a_j|m,n)P(f_j|e_{a_j}) \tag{11}$$

$$= \prod_{j=1}^{n} \sum_{a_j=0}^{m} P(a_j|m,n)P(f_j|e_{a_j}) \tag{12}$$

IBM1: posterior

Posterior

$$P(a_1^n|f_1^n, e_0^m) = \frac{P(f_1^n, a_1^n|e_0^m)}{P(f_1^n|e_0^m)} \tag{13}$$

Factorised

$$P(a_j|f_1^n, e_0^m) = \frac{P(a_j|m,n)P(f_j|e_{a_j})}{\sum_{i=0}^{m} P(i|m,n)P(f_j|e_i)} \tag{14}$$

MLE via EM

E-step:

$$\mathbb{E}[n(\mathsf{e} \to \mathsf{f}|A_1^n)] = \sum_{a_1=0}^{m} \cdots \sum_{a_n=0}^{m} P(a_1^n|f_1^n, e_0^m) n(\mathsf{e} \to \mathsf{f}|A_1^n) \tag{15}$$

$$= \sum_{a_1=0}^{m} \cdots \sum_{A_n=0}^{m} \prod_{j=1}^{n} P(a_j|f_1^n, e_0^m) \mathbb{1}_{\mathsf{e}}(e_{a_j}) \mathbb{1}_{\mathsf{f}}(f_j) \tag{16}$$

$$= \prod_{j=1}^{n} \sum_{i=0}^{m} P(A_j = i|f_1^n, e_0^m) \mathbb{1}_{\mathsf{e}}(e_i) \mathbb{1}_{\mathsf{f}}(f_j) \tag{17}$$

M-step:

$$\theta_{e,f} = \frac{\mathbb{E}[n(e \to f|A_1^n)]}{\sum_{f'} \mathbb{E}[n(e \to f'|A_1^n)]} \tag{18}$$

EM algorithm

Repeat until convergence to a local optimum

1. For each sentence pair
   1.1 compute posterior per alignment link
   1.2 accumulate fractional counts
2. Normalise counts for each English word

Content

Alignment distribution

Positional distribution

$P(A_j|M = m, N = n) = \mathrm{Cat}(A|\lambda_{j,m,n})$

- one distribution for each tuple (j, m, n)
- support must include length of longest English sentence
- extremely over-parameterised!

Alignment distribution

Positional distribution

$P(A_j|M = m, N = n) = \mathrm{Cat}(A|\lambda_{j,m,n})$

- one distribution for each tuple (j, m, n)
- support must include length of longest English sentence
- extremely over-parameterised!

Jump distribution [Vogel et al., 1996]

- define a jump function $\delta(a_j, j, m, n) = a_j - \left\lfloor j \frac{m}{n} \right\rfloor$
- $P(A_j|m, n) = \mathrm{Cat}(\Delta|\lambda)$
- Δ takes values from $-$longest to $+$longest

Content

Lexical alignment

Mixture models

IBM model 1

IBM model 2

Remarks

Note on terminology

Source/Target vs English/French

From an alignment model perspective all that matters is

- we condition on one language and generate the other
- in IBM models terminology, we condition on *English* and generate *French*

From a noisy channel perspective, where we want to translate a *source* sentence f_1^n into some *target* sentence e_1^m

- Bayes rule decomposes $p(e_1^m|f_1^n) \propto p(f_1^n|e_1^m)p(e_1^m)$
- train $p(e_1^m)$ and $p(f_1^n|e_1^m)$ independently
- **language model:** $p(e_1^m)$
- **alignment model:** $p(f_1^n|e_1^m)$
- note that the alignment model conditions on the target sentence (English) and generates the source sentence (French)

Limitations of IBM1-2

- too strong independence assumptions
- categorical parameterisation suffers from data sparsity
- EM suffers from local optima

Extensions

Fertility, distortion, and concepts [Brown et al., 1993]

Dirichlet priors and posterior inference [Mermer and Saraclar, 2011]

- + no NULL words [Schulz et al., 2016]
- + HMM and efficient sampler [Schulz and Aziz, 2016]

Log-linear distortion parameters and variational Bayes [Dyer et al., 2013]

First-order dependency (HMM) [Vogel et al., 1996]

- E-step requires dynamic programming [Baum and Petrie, 1966]

References I

L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37:1554–1563, 1966.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19 (2):263–311, June 1993. ISSN 0891-2017. URL `http://dl.acm.org/citation.cfm?id=972470.972474`.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

References II

Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N13-1073`.

Adrien Lardilleux and Yves Lepage. Sampling-based multilingual alignment. In *Proceedings of the International Conference RANLP-2009*, pages 214–218, Borovets, Bulgaria, September 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/R09-1040`.

References III

Coskun Mermer and Murat Saraclar. Bayesian word alignment for statistical machine translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 182–187, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-2032.

Philip Schulz and Wilker Aziz. Fast collocation-based bayesian hmm word alignment. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3146–3155, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL http://aclweb.org/anthology/C16-1296.

References IV

Philip Schulz, Wilker Aziz, and Khalil Sima'an. Word alignment without null words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 169–174, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://anthology.aclweb.org/P16-2028.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based word alignment in statistical translation. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 836–841, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics. doi: 10.3115/993268.993313. URL http://dx.doi.org/10.3115/993268.993313.