Lexical alignment: IBM models 1 and 2

MLE via EM for categorical distributions

Miguel Rios

April 7, 2019

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

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

Is there anything we could say about this language?

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:

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

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
- $\boxdot \iff$ cat

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

A few hypotheses:

- $\Box \iff$ dog
- $\boxdot \iff$ 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	el perro negro
the nice dog	el perro bonito
the black cat	el gato negro
a dog chasing a cat	un perro presiguiendo a un 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 | el perro negro

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

Data augmentation

Observations:

$$\text{the black dog} \mid \text{el perro negro}$$

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

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

Data augmentation

Observations:

the black dog | el perro negro

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

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

Data augmentation

Observations:

the black dog | el perro negro

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

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

Data augmentation

Observations:

$$\text{the black dog} \mid \text{el perro negro}$$

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

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

Data augmentation

Observations:

the black dog $\mid$ el perro negro

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

the black dog $\mid$ $(1, \text{the} \to \text{el})$ $(3, \text{dog} \to \text{perro})$ $(A_3, E_{A_3} \to F_3)$

Data augmentation

Observations:

the black dog | el perro negro

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

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

Data augmentation

Observations:

$$\text{the black dog} \mid \text{el perro negro}$$

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

$$\text{the black dog} \mid (1, \text{the} \rightarrow \text{el})\ (3, \text{dog} \rightarrow \text{perro})\ (2, \text{black} \rightarrow \text{negro})$$

Data augmentation

Observations:

the black dog $\mid$ el perro negro

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

the black dog $\mid$ $(1, \text{the} \rightarrow \text{el})$ $(3, \text{dog} \rightarrow \text{perro})$ $(2, \text{black} \rightarrow \text{negro})$

the black dog $\mid$ $(A_1, \text{the} \rightarrow \text{el})$ $(A_1, \text{the} \rightarrow \text{perro})$ $(A_1, \text{the} \rightarrow \text{negro})$

Data augmentation

Observations:

the black dog | el perro negro

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

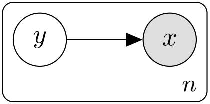
the black dog | $(1, \text{the} \to \text{el})$ $(3, \text{dog} \to \text{perro})$ $(2, \text{black} \to \text{negro})$

the black dog | $(A_1, \text{the} \to \text{el})$ $(A_1, \text{the} \to \text{perro})$ $(A_1, \text{the} \to \text{negro})$

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

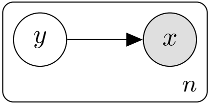
Content

Mixture models: generative story



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

Mixture models: generative story

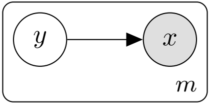


- c mixture components
- each defines a distribution over the same data space $\mathcal{X}$
- plus a distribution over components themselves

Generative story

1. select a mixture component $y \sim p(y)$
2. generate an observation from it $x \sim p(x|y)$

Mixture models: likelihood



Incomplete-data likelihood

$$p(x_1^m) = \prod_{i=1}^{m} p(x_i) \tag{1}$$

$$= \prod_{i=1}^{m} \sum_{y=1}^{c} \underbrace{p(x_i, y)}_{\text{complete-data likelihood}} \tag{2}$$

$$= \prod_{i=1}^{m} \sum_{y=1}^{c} p(y) p(x_i|y) \tag{3}$$

Interpretation

Missing data

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

Interpretation

Missing data

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

Inference: posterior distribution over possible y for each x

$$p(y|x) = \frac{p(y,x)}{\sum_{y'=1}^{c} p(y',x)} \tag{4}$$

$$= \frac{p(y)p(x|y)}{\sum_{y'=1}^{c} p(y')p(x|y')} \tag{5}$$

Non-identifiability

Different parameter settings, same distribution

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

y	$x = a$	$x = b$
1	0.2	0.8
2	0.7	0.3
$p(x)$	0.45	0.55

y	$x = a$	$x = b$
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 $p(y = 1) = p(y = 2) = 0.5$

y	$x = a$	$x = b$
1	0.2	0.8
2	0.7	0.3
$p(x)$	0.45	0.55

y	$x = a$	$x = b$
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

$$p(\mathcal{D}) = \prod_{i=1}^{m} p_\theta(x^{(i)})$$

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

$$p(\mathcal{D}) = \prod_{i=1}^{m} p_\theta(x^{(i)})$$

the score function is

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

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

$$p(\mathcal{D}) = \prod_{i=1}^{m} p_\theta(x^{(i)})$$

the score function is

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

then we choose

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

MLE for categorical: estimation from fully observed data

Suppose we have **complete data**

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

MLE for categorical: estimation from fully observed data

Suppose we have **complete data**

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

Then, for a **categorical distribution**

$$p(x|y) = \theta_{y,x}$$

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

MLE solution:

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

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 y happened with probability $p_\theta(y|x)$

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

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 $p(x|y) = \theta_{y,x}$
  y and x are categorical
  $0 \leq \theta_{y,x} \leq 1$ and $\sum_{x \in X} \theta_{y,x} = 1$

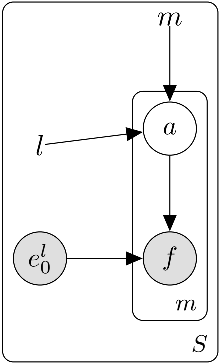
$$\theta_{y,x} = \frac{\mathbb{E}[n(y \to x|\mathcal{D}_{\text{completed}})]}{\sum_{x'} \mathbb{E}[n(y \to x'|\mathcal{D}_{\text{completed}})]} \quad (6)$$

$$= \frac{\sum_{i=1}^{m} \sum_{y'} p(y'|x^{(i)}) \mathbb{1}_y(y') \mathbb{1}_x(x^{(i)})}{\sum_{i=1}^{m} \sum_{x'} \sum_{y'} p(y'|x^{(i)}) \mathbb{1}_y(y') \mathbb{1}_{x'}(x^{(i)})} \quad (7)$$

$$= \frac{\sum_{i=1}^{m} p(y|x^{(i)}) \mathbb{1}_x(x^{(i)})}{\sum_{i=1}^{m} \sum_{x'} p(y|x^{(i)}) \mathbb{1}_{x'}(x^{(i)})} \quad (8)$$

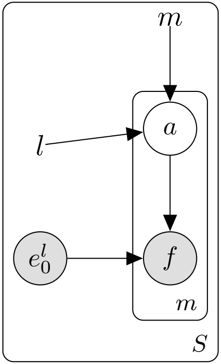
Content

IBM1: a constrained mixture model



Constrained mixture model

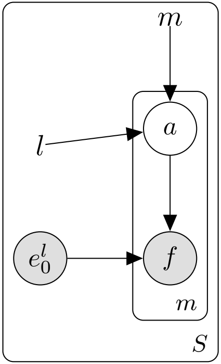
IBM1: a constrained mixture model



Constrained mixture model

- mixture components are English words

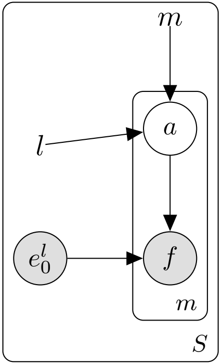
IBM1: a constrained mixture model



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 NULL component
- j ranges over French words and i over English words

Parameterisation

Alignment distribution: uniform

$$p(a|l,m) = \frac{1}{l+1} \tag{9}$$

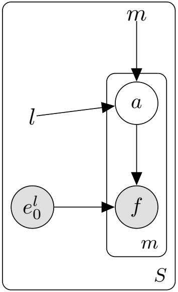
Lexical distribution: categorical

$$p(f|e) = \mathrm{Cat}(f|\theta_e) \tag{10}$$

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

IBM1: incomplete-data likelihood

Incomplete-data likelihood



$$p(f_1^m|e_0^l) = \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} p(f_1^m, a_1^m|e_{a_j}) \tag{11}$$

$$= \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{n} p(a_j|l,m)p(f_j|e_{a_j}) \tag{12}$$

$$= \prod_{j=1}^{n} \sum_{a_j=0}^{l} p(a_j|l,m)p(f_j|e_{a_j}) \tag{13}$$

IBM1: posterior

Posterior

$$p(a_1^m|f_1^m, e_0^l) = \frac{p(f_1^m, a_1^m|e_0^l)}{p(f_1^m|e_0^l)} \tag{14}$$

Factorised

$$p(a_j|f_1^m, e_0^l) = \frac{p(a_j|l, m)p(f_j|e_{a_j})}{\sum_{i=0}^{l} p(i|l, m)p(f_j|e_i)} \tag{15}$$

MLE via EM

E-step:

$$\mathbb{E}[n(\mathsf{e} \to \mathsf{f}|a_1^m)] = \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} p(a_1^m|f_1^m, e_0^l) n(\mathsf{e} \to \mathsf{f}|A_1^m) \tag{16}$$

$$= \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} p(a_j|f_1^m, e_0^l) \mathbb{1}_{\mathsf{e}}(e_{a_j}) \mathbb{1}_{\mathsf{f}}(f_j) \tag{17}$$

$$= \prod_{j=1}^{m} \sum_{i=0}^{l} p(a_j = i|f_1^m, e_0^l) \mathbb{1}_{\mathsf{e}}(e_i) \mathbb{1}_{\mathsf{f}}(f_j) \tag{18}$$

M-step:

$$\theta_{e,f} = \frac{\mathbb{E}[n(e \to f|a_1^m)]}{\sum_{f'} \mathbb{E}[n(e \to f'|a_1^m)]} \tag{19}$$

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

$p(a_j|l,m) = \mathrm{Cat}(a|\lambda_{j,l,m})$

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

Alignment distribution

Positional distribution

$p(a_j|l,m) = \mathrm{Cat}(a|\lambda_{j,l,m})$

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

Jump distribution [Vogel et al., 1996]

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

Content

Decoding

- Pick the alignment that has the highest posterior probability.
- Assumption conditional independence of alignment links
  Maximising the probability of an alignment factorises over individual alignment links.
- $arg\, maxp(a_1^m \mid f_1^m, e_0^l)$

Content

Note on terminology: source/target vs French/English

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^l

- Bayes rule decomposes $p(e_1^l|f_1^n) \propto p(f_1^n|e_1^l)p(e_1^l)$
- train $p(e_1^l)$ and $p(f_1^n|e_1^l)$ independently
- **language model:** $p(e_1^l)$
- **alignment model:** $p(f_1^n|e_1^l)$
- 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.