在多语言BERT中寻找语言线索以提高跨语言泛化

Chi-LiangLiu*song - yuanHsu*Chuang Chung-Yi Li Hung-Yi Lee
国立台湾大学电气工程与计算机学院{liangtaiwan1230, sivia89024,
tlkagkb93901106}@gmail.com
{b05901033, r07942080} @ntu.edu.tw

摘要

多语言BERT (m-BERT)中的标记嵌入包含语言和语义信息。我们发现，一种语言的表示可以通过简单地平均该语言的标记的嵌入来获得。给定这种语言表示，我们通过操纵标记嵌入来控制多语言BERT的输出语言，从而实现无监督标记翻译。在此基础上，我们进一步提出了一种计算成本低但有效的方法来提高m-BERT的跨语言能力。

1 介绍

多语言BERT (m-BERT) (Devlin等人，2019)已经证明了它在各种任务的跨语言迁移中的优势(Conneau等人，2018;Wu和Dredze, 2019;许等人，2019;Pires等人，2019);这归功于其内部表征的跨语言对齐，其中来自不同语言的语义相似或功能相似的单词用相似的嵌入表示(Cao等人，2020;Liu et al.，2020)。

语言信息是如何嵌入到m-BERT中的?答案可能比想象中更直接。我们发现一种语言的平均token嵌入很好地代表了该语言。为了验证这一观察结果，我们表明，如果将一个英语句子输入到m-BERT中，然后将其嵌入在嵌入空间中的特定方向上移动，则m-BERT会以语义接近输入句子的另一种语言输出一个句子。

在证明了m-BERT嵌入中存在语言信息并有了一种简单的提取方法之后，我们消除了嵌入中这些特定于语言的变化，并证明了这是一种提高零-的实用方法

m-BERT在下游任务中的跨语可转移性。

在文献中，一些人试图改善预训练的m-BERT的跨语言对齐。例如，Cao等人(2020)提出在小型并行数据集上微调m-BERT, Libovický等人(2020)提出逐语言对嵌入语言进行零中心，以实现语言中性性并展示检索任务的进展，这些都是在无监督场景下进行的。我们的工作与(Libovický et al.，2020;Gonen et al.，2020)的和与其方法有相似之处。与它们相比，我们讨论了更多的语言表示应用。

这项工作的贡献可以总结为以下几点:

- m-BERT中的语言信息可以用特定语言的所有标记嵌入的平均值来表示。这可以通过无监督的token翻译来验证。
- m-BERT在下游任务中的跨语言可转移性可以通过操纵标记嵌入来提高。

2 语言表示

假设我们有n种语言，用{L1, L2，…，Ln}及其对应的语料库。

上下文依赖的表示给定一个输入序列x和标记索引i，我们用h_{lx_i}表示层l中的隐藏表示。

语言均值给定一个语言L和它对应的由一组句子x组成的语料库C，我们表示层L的语言均值为

$$\boldsymbol{R}_L^l = \mathop{\mathbb{E}}_{x_i \in C}\left[\boldsymbol{h}_{x_i}^l\right],$$

其中表示语料库中所有标记嵌入的均值。我们假设语言

*Equal Contribution

Mean包含特定于语言的信息，但没有语义信息。

虽然这个关于语言表示的假设在这里看起来很幼稚，但在实验中，我们表明RLl很好地表示了标记嵌入中的语言信息。对于每一种语言L，每一层L都有一个特定于语言的表示RLl。因为我们不知道L RLl最能代表语言L，所以L在以下算法中是一个超参数。

为了消除特定于语言的信息，我们只需减去语言平均RLl$_k$
从每个token嵌入h$_{lx_i}$，从而将token嵌入移动到一个与语言无关的联合空间。与语言无关的隐藏表示h⊠l $_{xican}$可以写成

$$\hat{\boldsymbol{h}}^l_{x_i} = \boldsymbol{h}^l_{x_i} - \boldsymbol{R}^l_{L_k}, \quad (1)$$

其中h$_{lx_i}$是从Lk中的token xi中提取的。

Mean Difference Shift (MDS)除了消除语言信息外，我们还可以将L1空间中的嵌入移动到L2空间:这相当于无监督令牌翻译。也就是说，给定*我*在英语中的嵌入，我们修改嵌入，使其被m-BERT解释为我在中文中的嵌入。

我们将句子L1输入到m-BERT中，并在第1层提取每个标记的嵌入。然后我们从嵌入中减去RLl，如(1)中所示
移动L1的信息，然后将RLl加到$_2$
将嵌入转移到L2空间。形式上，我们将L1中的token embeddingh$_{lx_i}$修改为embedding h ~ l $_{xiin}$ L2 as

$$\tilde{\boldsymbol{h}}^l_{x_i} = \boldsymbol{h}^l_{x_i} + \boldsymbol{R}^l_{L_2} - \boldsymbol{R}^l_{L_1}. \quad (2)$$

3 无监督Token翻译

在本节中，我们展示了嵌入空间中隐含的特定于语言的信息可以从语义嵌入中解耦出来。我们使用MDS将L1中的句子输入到m-BERT中，*并将其翻译成*L2中的句子。

3.1 设置

MDS的公式对式(2)稍作修改:h ~ lxi = hlxi + α rll2−RLl1，其中α为超参数;我们将在实验结果中看到它的影响。给定输入，在特定层l修改to-ken嵌入。

第(l +1)层将修改后的嵌入作为输入，最后一层生成token序列。本实验中的句子来自XNLI测试集，该测试集包含15种语言，包括斯瓦希里语和乌尔都语等低资源语言。

3.2 评价指标

我们使用两种不同的指标来定量分析无监督token翻译的结果。
BLEU-1分数该指标在不考虑转换序列流畅性的情况下衡量了翻译质量。

转化率除了翻译质量，我们还计算了*转化率*:从源语言转换到目标语言的token的百分比，定义为

$$\text{conversion rate} = \frac{\#\text{ of } y \in (V_t - V_s)}{\#\text{ of } y - \#\text{ of } y \in V_s \cap V_t},$$

其中y是模型的输出token, Vs和Vt是源语言和目标语言的token集。由于两个词汇表共享的token未被考虑在内，因此它们被排除在分子和分母项之外。

3.3 结果

令人惊讶的是，我们能够通过应用MDS在给定L1输入的语言L2中产生预测的标记;许多预测的标记是L1中输入标记的标记级翻译，即使是低资源语言。示例输出如附录A所示。

表1为量化结果。首先，尽管翻译结果不及现有的无监督翻译方法(Kim et al., 2018)，但它构成了强有力的证据，表明我们可以使用MDS在令牌嵌入空间中操纵特定于语言的信息，然后诱导m-BERT从一种语言切换到另一种语言。其次，我们观察到，随着α的增加，模型将更多的token转换为目标语言L2，并且永远不会解码不同时属于L1和L2的token。给定负的α，模型总是解码属于L1的词例。这表明在嵌入空间中，与语言相关的方向是唯一的。我们在附录A中提供了进一步的α分析。

表1:使用第10层BERT的定量无监督令牌翻译结果

	en→de	en→fr	en→ur	en→sw	en→zh	en→el	de→en	fr→en	ur→en	sw→en	zh→en	el→en
BLEU-1 (α=1)	7.53	8.53	5.56	7.96	15.25	7.88	7.34	9.08	5.52	6.34	4.37	6.54
BLEU-1 (α=2)	8.03	10.24	6.31	7.23	21.51	14.91	7.48	8.52	6.23	7.48	5.38	6.65
BLEU-1 (α=3)	12.35	10.65	5.35	7.16	15.95	19.13	6.29	12.27	5.74	6.45	6.17	4.73
Conversion rate (α=1)	40.2	41.7	61.1	15.3	47.8	62.1	45.2	49.6	29.9	14.7	23.9	30.2
Conversion rate (α=2)	74.8	75.7	99.4	97.4	90.0	99.1	67.3	601	83.0	65.6	60.8	97.9
Conversion rate (α=3)	95.2	96.3	99.8	100	99.5	100	79.5	73.1	96.6	93.6	91.4	99.7

表2:基于第8层BERT的Tatoeba句子检索

Method	de	es	ar	el	fr	hi
Original	75.4	64.1	24.5	29.8	64.3	**34.8**
MUSE	1.3	0.2	0.3	0.5	23.8	0.2
Zero-mean	73.5	61.8	23.5	29.4	63.7	29.9
MDS	**76.8**	**67.5**	**29.1**	**30.6**	**67.0**	31.4
	ru	vi	th	tr	zh	
Original	**63.6**	**61.0**	13.7	32.9	68.6	
MUSE	0.2	0.2	0.4	0.1	0.2	
Zero-mean	59.6	51.2	13.7	32.8	64.1	
MDS	59.4	51.5	**17.5**	**36.8**	**69.2**	

表3:基于第8层BERT的BUCC2018开发集和测试集句子检索

	Method	de	fr	ru	zh
Dev	Original	75.62	72.07	68.59	66.04
	Zero-mean	71.10	70.51	65.92	59.91
	MUSE	13.68	57.11	40.49	16.98
	MDS	**76.91**	**73.45**	**71.60**	**66.91**
Test	Original	63.22	62.47	11.65	50.47
	Zero-mean	59.59	59.25	10.40	45.42
	MUSE	59.00	11.30	6.03	2.03
	MDS	**65.76**	**63.95**	**12.36**	**52.45**

4 跨语言句子检索

从两种语言之间的可比语料库中提取平行句是评估跨语言嵌入的常用方法(Hu等人，2020;Zweigenbaum 等 人 ， 2017;Artetxe 和 Schwenk, 2018)。在本节中，我们使用对句子级检索任务的评估来演示MDS实现更好的跨语言对齐。

4.1 任务

我们评估了MDS和零均值在BUCC2018和Tatoeba两个句子检索任务上的效果。我们使用句子中所有token嵌入的均值向量作为句子嵌入，余弦相似度作为距离度量。从BERT编码器的特定层提取Token嵌入，并将在整个数据集上预先计算的MDS或零均值移位直接应用于提取的嵌入。

4.2 MDS与零均值

虽然在句子检索任务中应用MDS或零均值乍一看似乎相似，但它们之间存在微妙的差异。假设句子嵌入v1∈L1和v2∈L2，并且这两个不同语言的句子具有相同的语义。假设存在真实的语言表示RL *和RL *完全符合
从嵌入中消除语言，使得v1−RL*[1]= v2−RL*2。通过平均得到的语言表征RL_1和RL_2是真实表征的近似值，δ1和δ2是真实与近似语言表征的差值。

$$\begin{aligned} &\boldsymbol{v}_1 - \boldsymbol{R}_{L_1} \neq \boldsymbol{v}_2 - \boldsymbol{R}_{L_2} \\ \rightarrow &\boldsymbol{v}_1 - \boldsymbol{R}_{L_1} - \boldsymbol{\delta}_1 = \boldsymbol{v}_2 - \boldsymbol{R}_{L_2} - \boldsymbol{\delta}_2 \end{aligned}$$

则v1和v2的后mds和后零均值余弦相似度为

$$\begin{aligned} \cos_{\text{MDS}}(\boldsymbol{v}_1, \boldsymbol{v}_2) &= \frac{(\boldsymbol{v}_1 - \boldsymbol{R}_{L_1} + \boldsymbol{R}_{L_2}) \cdot \boldsymbol{v}_2}{|\boldsymbol{v}_1 - \boldsymbol{R}_{L_1} + \boldsymbol{R}_{L_2}||\boldsymbol{v}_2|} \\ &= \frac{(\boldsymbol{v}_2 + \boldsymbol{\delta}) \cdot \boldsymbol{v}_2}{|\boldsymbol{v}_2 + \boldsymbol{\delta}||\boldsymbol{v}_2|}, \text{ where } \boldsymbol{\delta} = \boldsymbol{\delta}_1 - \boldsymbol{\delta}_2 \\ \cos_{\text{zero-mean}}(\boldsymbol{v}_1, \boldsymbol{v}_2) &= \frac{(\boldsymbol{v}_1 - \boldsymbol{R}_{L_1}) \cdot (\boldsymbol{v}_2 - \boldsymbol{R}_{L_2})}{|\boldsymbol{v}_1 - \boldsymbol{R}_{L_1}||\boldsymbol{v}_2 - \boldsymbol{R}_{L_2}|} \\ &= \frac{(\boldsymbol{v}_2 - \boldsymbol{R}^*_{L_2} + \boldsymbol{\delta}_1)(\boldsymbol{v}_2 - \boldsymbol{R}^*_{L_2} + \boldsymbol{\delta}_2)}{|\boldsymbol{v}_2 - \boldsymbol{R}^*_{L_2} + \boldsymbol{\delta}_1||\boldsymbol{v}_2 - \boldsymbol{R}^*_{L_2} + \boldsymbol{\delta}_2|}. \end{aligned}$$

这表明，当|v2| > |v2−RL_2| > max (|δ 1，|δ 2|，|δ|)时，零均值对近似误差更敏感。3在实验中进一步验证了两种方法的差异性。

4.3 结果

句子检索结果如表3和表2所示。在BUCC2018开发集和测试集上，mds移位嵌入在所有语言上始终产生更高的准确性，零均值嵌入比什么都不做更糟糕。在

[1] Unknown to us

[3] Superscript [1] ignored here for simplicity

This is very possible. Because v_2 is in L_2, it may have the same direction as R_{L_2}.

表4:词性标注结果

Method	ar	bg	de	el	es	fr	hi	ru	th	tr	ur	vi	zh	Average
Original	53.8	85.4	86.2	81.1	86.1	42.9	66.8	85.5	41.7	68.6	56.3	**53.8**	61.8	66.9
Zero-mean	**54.3**	86.1	**86.6**	**81.8**	86.6	43.7	68.1	**86.5**	41.6	**69.7**	56.6	53.4	62.5	67.5
MDS	54.2	**86.4**	86.5	81.5	**86.8**	**43.9**	**68.9**	86.4	**44.2**	69.4	**57.1**	52.4	**63.0**	**67.8**

表5:依存分析结果。数字被标记为附件得分(LAS)。

Method	ar	bg	de	el	es	fr	hi	ru	th	tr	ur	vi	zh	Average
Original	28.2	70.7	**74.0**	**71.6**	72.1	74.8	35.3	69.0	30.8	32.9	28.3	**37.8**	**35.4**	50.8
Zero-mean	**28.2**	**71.0**	73.4	71.4	72.2	**75.7**	36.3	**69.3**	**32.5**	**34.6**	28.6	37.0	35.2	**51.2**
MDS	28.0	70.8	73.7	71.1	**72.2**	75.3	**36.5**	68.8	30.4	34.2	**29.0**	35.6	35.0	50.8

在Tatoeba测试集上，除了印地语、俄语和越南语外，MDS嵌入在大多数语言中也是最好的。我们还尝试使用旋转矩阵来对齐嵌入(MUSE，(Lample等人，2017))，但发现无监督对齐方法在BERT上不起作用。

5 跨语言迁移

5.1 设置

在零概率跨语言迁移学习中，m-BERT在源语言上进行了微调，在接下来的实验中，源语言是英语，在微调过程中从未见过的语言上进行了测试。对于每种语言，我们使用了来自维基百科文档的大约5M个标记来计算语言表示。

零均值在微调期间，我们对源语言的token嵌入应用零均值，并在训练期间将修改后的嵌入转发到其余层。在测试期间，我们将目标语言数据输入微调模型，并将零均值应用于第l层的嵌入。使用预训练模型从维基百科数据中提取语言向量均值。

在这种方法中，我们没有在训练期间修改嵌入。在测试过程中，我们将MDS应用于第1层的嵌入。使用微调模型从维基百科数据中提取平均差向量。

5.2 任务

为了证明所提出的方法提高了跨语言零学习性能，我们在两个任务上进行了实验:词性(POS)标注和依赖关系分析。

对于POS，我们使用了90种语言的Universal Dependencies v2.5 (Nivre等人，2020)树库。每个单词分配17个通用POS标签中的一个。该模型在英语上进行训练，并在其他13种语言上进行测试。

对于依赖项解析，数据集和跨语言传输设置与POS标记完全相同。

5.3 结果

表4和表5分别比较了基线和我们的方法在POS标记和依赖项解析方面的结果。对于POS标注，零均值和MDS都提高了跨语言测试集上的性能，只有少数例外。MDS在泰语中没有帮助(th)，并且两种方法都没有改善越南语(vi)。对于依赖关系解析，与之前的观察相反，我们发现只有零均值在基线上得到改善，而MDS没有。由于影响依赖解析的语言因素(例如头部方向参数)比POS更多，我们认为需要更多的分析来解释依赖解析的性能。

6 结论

本文研究了m-BERT嵌入中特定语言信息的存在性，并通过操纵特定语言信息实现了无监督的token翻译。所提出的方法进一步被证明在改善跨语言嵌入对齐和跨语言迁移学习方面是有效的。我们将在更多下游任务上进一步探索所提出的方法。

参考文献

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representa-tions. In *International Conference on Learning Rep-resentations*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceed-ings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language under-standing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Associ-ation for Computational Linguistics.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not Greek to mBERT: Induc-ing word-level translations from multilingual BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 45–56, Online. Association for Compu-tational Linguistics.

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. In *Proceedings of the 2019 Conference on Empirical Methods in Natu-ral Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5933–5940, Hong Kong, China. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task bench-mark for evaluating cross-lingual generalization.

Yunsu Kim, Jiahui Geng, and Hermann Ney. 2018. Improving unsupervised word-by-word translation with language model and denoising autoencoder. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 862–868.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc' Aurelio Ranzato. 2017. Unsupervised ma-chine translation using monolingual corpora only. *arXiv preprint arXiv: 1711.00043*.

Jindˇrich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained mul-tilingual representations.

Chi-Liang Liu, Tsung-Yuan Hsu, Yung-Sung Chuang, and Hung yi Lee. 2020. What makes multilingual BERT multilingual? In *arXiv*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Gin-ter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4034–4043, Mar-seille, France. European Language Resources Association.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996– 5001, Florence, Italy. Association for Computa-tional Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natu-ral Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Com-putational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable cor-pora. In *Proceedings of the 10th Workshop on Build-ing and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

无监督Token翻译中的α分析

表6:token集的大小和英语token集与另一种语言token集交集的大小

	en	de	fr	el	zh	ur	sw
$\lvert V_{\text{lang}}\rvert$	9140	9212	8552	3189	3866	4085	5609
$\lvert V_{en} \cap V_{\text{lang}}\rvert$	9140	3230	3911	1696	1325	1549	2970

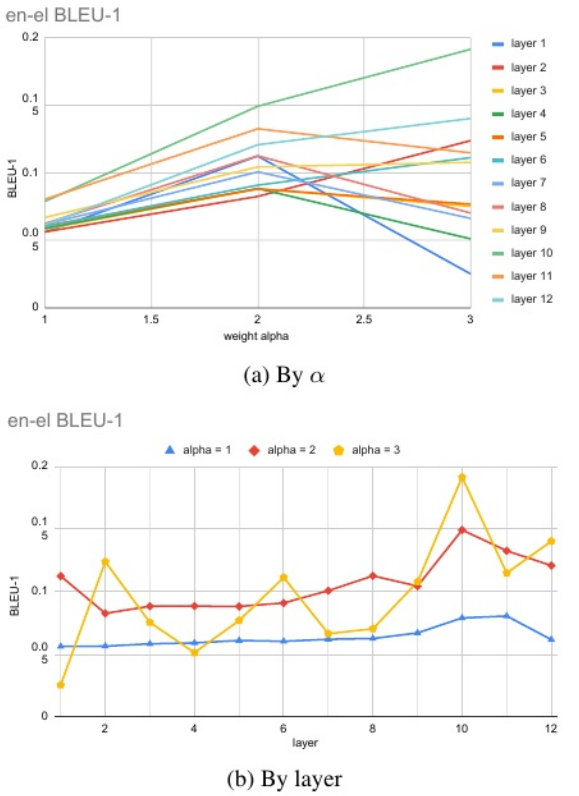

图1:不同α、不同层的无监督*en*) *el*令牌翻译BLEU-1的变化方向 -

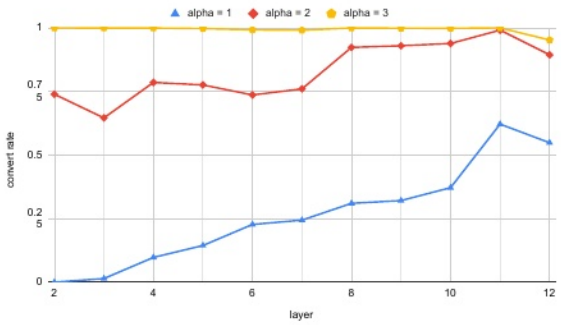

图2:MDS在不同层上给定不同α的*en*) *el*数据的转化率

我们在图1和图2*中*给出了一个示例，以显示转化率和BLEU-1分数随不同α权重和不同层的变化情况。

尽管权重增加对BLEU-1的影响是混合的，但在最后几层(10层或11层)中，当α设置为3.0时，大多数语言的BLEU-1显著上升(表1中最佳层行也显示了这一点)。这表明最后几层更适合于分离语言特定表示，这与文献中的观察一致，即最后几层包含更多用于预测被屏蔽词的语言特定信息(Pires et al.， 2019)。

表7:随机样本的无监督token翻译(MDS，第10层)

Input (en)	The girl that can help me is all the way across town. There is no one who can help me.
Ground truth (zh)	能帮助我的女孩在小镇的另一边。没有人能帮助我。。
en→zh, $\alpha = 1$	. 孩，can 来我是all the way across 市。。There 是无人人can help 我。
en→zh, $\alpha = 2$	. 孩的的家我是这个人的市。。他是他人人的到我。
en→zh, $\alpha = 3$	。，的的的他是的个的的，。：他是他人，的。他。
Ground truth (fr)	La fille qui peut m'aider est à l'autre bout de la ville. Il n'y a personne qui pourrait m'aider.
en→fr, $\alpha = 1$	. girl qui can help me est all la way across town . . There est no one qui can help me .
en→fr, $\alpha = 2$	. girl qui de help me est all la way dans , . . Il est de seul qui pour aid me .
en→fr, $\alpha = 3$	, , , de , me , all la , , , , , n n n n , , , , ,