Chapter 1: See Other File

Chapter 2: See Other File

Chapter 3: See Other File

Chapter 4: See Other File

Chapter 5: See Other File

Chapter 6: Expert *Annotation* and *Evaluation*

We introduce a new computational task, adaptation, where the gold standard is subjective and all-important, thereby requiring authoritative **experts**, rather than the anonymous **crowd** (Chapter **??**). Machine translation usually translates words literally; however, this does not necessarily apply in a cultural context as certain named entities may be relevant in one culture but not another. We propose three methods—two computational and one expert-driven—to find such named entities across American and German culture. **Annotation** for the human method requires specialized knowledge: familiarity with German or American culture. Furthermore, **evaluation** requires knowledge of *both* cultures. This chapter explores the use of experts for an **annotation** task. Chapter **??** will use them for **generation**.

6.1 When Translation Misses the Mark

Imagine reading a translation from German, "I saw Merkel eating a Berliner from Dietsch on the ICE". This sentence is opaque without cultural context.

An extreme cultural *adaptation* for an American audience could render the sentence as "I saw Biden eating a Boston Cream from Dunkin' Donuts on the Acela", elucidating that Merkel is in a similar political post to Biden; that Dietsch (like

Top Adaptations for Bill Gates:

WikiData	**3CosAdd**	**Human**
F. Zeppelin	congstar	A. Bechtolsheim
Günther Jauch	Alnatura	Dietmar Hopp
N. Harnoncourt	GMX	Carl Benz

Table 6.1: WikiData and unsupervised embeddings (**3CosAdd**) generate adaptations of an entity, such as Bill Gates. Human adaptations are gathered for evaluation. American and German entities are color coded.

Dunkin' Donuts) is a mid-range purveyor of baked goods; both Berliners and Boston Creams are filled, sweet pastries named after a city; and ICE and Acela are slightly ritzier high-speed trains. Human translators make this adaptation when it is appropriate to the translation (Gengshen, 2003).

Because adaptation is understudied, we leave the full translation task, which requires **generation**, to future work. Instead, we focus on the task of cultural adaptation, akin to **annotation**, of entities: given an entity in a source, what is the corresponding entity in English? Most Americans would not recognize Christian Drosten, but the most efficient explanation to an American would be to say that he is the "German Anthony Fauci" (Loh). We provide top adaptations suggested by algorithms and humans for another American involved with the pandemic response, Bill Gates, in Table 6.1.

Can machines reliably find these analogs with minimal supervision? We generate these adaptations with structured knowledge bases (Section 6.3) and word embeddings (Section 6.4). We elicit human adaptations (Section 6.5) to evaluate whether our automatic adaptations are plausible (Section 6.5.3).

6.2 Wer ist Bill Gates?

We define cultural adaptation and motivate its application for tasks like creating culturally-centered training data for QA. (Vinay and Darbelnet, 1995) define adaptation as translation in which the relationship not the literal meaning between the receiver and the content needs to be recreated.

You could formulate our task as a traditional analogy Drosten::Germany as Fauci::United States (Turney, 2008; Gladkova et al., 2016), but despite this superficial resemblance (explored in Section 6.4), traditional approaches to analogy ignore the influence of culture and are typically *within* a language. Hence, analogies are tightly bound with culture; humans struggle with analogies outside their culture (Freedle, 2003).

We can use this task to identify named entities (Kasai et al., 2019; Arora et al., 2019; Jain et al., 2019) and for understanding other cultures (Katan and Taibi, 2004).

6.2.1 ... and why Bill Gates?

This task requires a list of named entities adaptable to other cultures. Our entities come from two sources: a subset of the top 500 most visited German/English Wikipedia pages and the non-official characterization list (Veale, 2016, NOC), "a source of stereotypical knowledge regarding popular culture, famous people (real and fictional) and their trade-mark qualities, behaviours and settings". Wikipedia contains a plethora of singers and actors; we filter the top 500 pages to avoid a pop culture skew.[1] We additionally select all Germans and a subset of Americans from the Veale NOC list as it is human-curated, verified, and contains a broader historical period than popular Wikipedia pages. Like other semantic relationships (Boyd-Graber et al., 2006), this is not symmetric. Thus, we adapt entities in both directions; while Berlin is the German Washington, DC, there is less consensus on what is the American Berlin, as Berlin is both the capital, a tech hub, and a film hub. A full list of our entities is provided in Appendix **??**.

6.3 Adaptation from a Knowledge Base

We first adapt entities with a knowledge base. We use WikiData (Vrandečić and Krötzsch, 2014), a structured, human-annotated representation of Wikipedia entities that is actively developed. This resource is well-suited to the task as features are standardized both within and across languages.

[1] We discuss the applicability of using Wikipedia (i.e., what proportion of the English Wikipedia is visited from the United States) in Appendix **??**.

Many knowledge bases explicitly encode the nationality of individuals, places, and creative works. Entities in the knowledge base are a discrete sparse vector, where most dimensions are unknown or not applicable (e.g., a building does not have a spouse). For example, Angela Merkel is a human (instance of), German (country of citizenship), politician (occupation), Rotarian (member of), Lutheran (religion), 1.65 meters tall (height), and has a PhD (academic degree). How would we find the "most similar" American adaptation to Angela Merkel? Intuitively, we should find someone whose nationality is American.

Some issues immediately present themselves; contemporary entities will have more non-zero entries than older entities. Some characteristics are more important than others: matching unique attributes like "worked as journalist" is more important than matching "is human".

Each entity in WikiData has "properties", which we can think about as the dimension of a sparse vector and "values" that those properties can take on. For example, Merkel has the properties "occupation" and "academic degree". *Values* for those properties are that her "occupation" is "politician" and her "academic degree" is a "doctorate". To match entities across cultures, we focus on matching properties rather than values; many of the values are more relevant inside a culture. For example, we cannot find American politicians who belong to the Christian Democratic Union, but we can find politicians who have an academic degree and a dissertation title.

As a toy example, if Beethoven, Merkel, and Bach all have only two *properties*: Beethoven has an "occupation" and "genre", Merkel has an "Erdős number" and "political party", and Bach has a "occupation" and "genre", then Beethoven and Bach has a distance of zero and are the closest entities while Merkel has a distance of two since {"Erdős number", "political party"} is two away from {"occupation", "genre"}.

First, we bifurcate WikiData into two sets: an American set $\mathcal{A}$ for items which contain the *value* "United States of America" and a German set $\mathcal{D}$ for those with German values.[2] This is a liberal approximation, but it successfully excludes roughly seven out of the eight million items in WikiData. Then we explore the *properties* from WikiData. We create entity vectors with dimensions corresponding to frequently-occurring properties.

The *properties* are discrete and categorical; Merkel either has an "occupation" or she does not. Each entity then has a sparse vector. We calculate the similarity of the vectors with Faiss's L_2 distance (Johnson et al., 2017) and for each vector in $\mathcal{A}$ find the closest vector in $\mathcal{D}$ and *vice versa*.

So who is the American Angela Merkel? One possible answer is Woodrow Wilson, a member of a "political party", who had a "doctoral advisor" and a "religion", and ended up with "awards". This answer may be unsatisfying as it was Barack Obama who sat across from Merkel for nearly a decade. To capture these more nuanced similarities, we turn to large text corpora in Section 6.4.

[2]While the geopolitical definition of American is straightforward, the German nation state is more nuanced (Schulze, 1991). Following Green (2003), we adopt members of the Zollverein or the German Confederation as "German" *as well as their predecessor and successor states*. This approach is a more inclusive (Großdeutschland) definition of"German" culture.

6.4 An Alternate Embedding Approach

While the classic NLP vector example (Mikolov et al., 2013c) isn't as magical as initially claimed (Rogers et al., 2017), it provides useful intuition. We can use the intuitions of the cliché:

$$\overrightarrow{\text{King}} - \overrightarrow{\text{Man}} + \overrightarrow{\text{Woman}} = \overrightarrow{\text{Queen}} \tag{6.1}$$

to adapt between languages.

This, however, requires relevant embeddings. First, we use the entire Wikipedia in English and German, preprocessed using Moses (Koehn et al., 2007). We follow Mikolov et al. (2013b) and use named entity recognition (Honnibal et al., 2020) to tokenize entities such as Barack_Obama.

We use word2vec (Mikolov et al., 2013b), rather than FastText (Bojanowski et al., 2016), as we do not want orthography to influence the similarity of entities. Angela Merkel in English and in German have quite different neighbors, and we intend to keep it that way by preserving the distinction between languages.

However, the standard word2vec model assumes a single monolingual embedding space. We use unsupervised Vecmap (Artetxe et al., 2018), a leading tool for creating cross-lingual word embeddings, to build bilingual word embeddings. We propose two approaches for adaptation.

3CosAdd We follow the word analogy approach of 3CosAdd[3] (Levy and Goldberg, 2014; Köper et al., 2016). American→German adaptation takes the source

[3]We experiment with 3CosMul as well but found 3CosAdd generally more robust.

entity's (v) embedding in the English vector space and looks for its adaptation (u^*) based on embeddings in the German space. This is like the word analogy task, i.e., what entity has the role in the German culture as v does in American culture. As an example, Merkel has a similar role in the German culture as Biden. Formally, the adaptation of the English entity v into German is

$$\vec{a} \equiv \text{avg}\left(\overrightarrow{E^{en}}_{\text{United\_States}}, \overrightarrow{E^{de}}_{\text{USA}}\right) \tag{6.2}$$

$$\vec{d} \equiv \text{avg}\left(\overrightarrow{E^{en}}_{\text{Germany}}, \overrightarrow{E^{de}}_{\text{Deutschland}}\right) \tag{6.3}$$

$$u^* = \arg\max_{u \in V^{de}} \text{sim}\left(\overrightarrow{E_u^{de}}, \overrightarrow{E_v^{en}} - \vec{a} + \vec{d}\right), \tag{6.4}$$

where $\overrightarrow{E_w^l}$ is the embedding of word w in language l, V^{de} is the German vocabulary and sim is the cosine similarity. The American anchor word $\vec{a}$ and German anchor $\vec{d}$ represent the American and German cultures.[4] We average the English and German embeddings of the individual word types for robust anchor vectors. In standard analogies, as in Equation 6.1, the $\vec{a}$ and $\vec{d}$ vectors are different for each test pair; here they are the same for each example, as we always are pivoting between the two cultures.

Learned adaptation To eliminate the need for manual anchor selection for both cultures, our second approach learns the adaptation as a linear transformation of source embeddings to the target culture given a few adaptation examples. Specifically, we use the human adaptations sourced for the Wikipedia entities as training for the Veale NOC ones. We follow the work of Mikolov et al. (2013a)

[4] USA refers to the United States in German. Der Spiegel, the largest newspaper, calls their US section USA.

and learn a transformation matrix $\mathbf{W}_{en \to de}$ for American→German by minimizing the L_2 distance of $\mathbf{W}_{en \to de} \overrightarrow{E}^{en}_{v_i}$ and $\overrightarrow{E}^{de}_{u_i}$ over gold adaptation $v_i, u_{i}{}_{i=1}^{n}$ entity pairs. The adaptation of a source entity v is $u^* = \mathbf{W}_{en \to de} \overrightarrow{E}^{en}_{v}$. Likewise, we learn the reverse mapping $\mathbf{W}_{de \to en}$ for German→American adaptation. This requires supervised training data—but not much (Conneau et al., 2017)—which we collect in Section 6.5.

6.5 Comparing Automation to Human Judgment

The computational methods can **generate** entities at scale, but humans have to **evaluate** their relevance.

6.5.1 Adaptation by Locals

Since quality control is difficult for **generation** and complicated **annotation** (Peskov et al., 2019; Karpinska et al., 2021), we need users who will answer the task accurately. We recruit five American citizens educated at American universities and five German citizens educated at German ones that are appropriately qualified **experts** for this task. These human **annotations** serve as a gold standard against which we can compare our **automatic** approaches. To improve the user experience, we create an interface (Figure 6.2) that provides a brief summary of each source entity from Wikipedia and asks the users to select a target adaptation that autocompletes Wikipedia page titles (all entities; targets are not limited to the lists in Section 6.2) in a text box *a la* answer selection in Wallace et al. (2019).

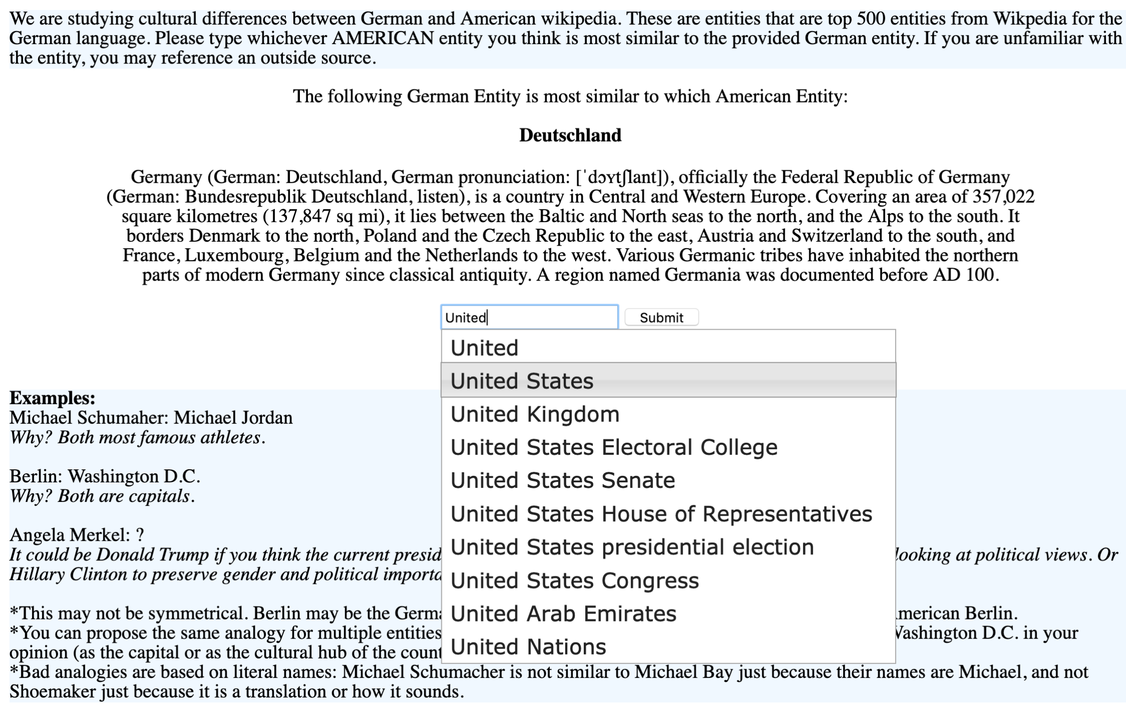

Figure 6.1: Our interface provides users with information about the entity and asks them to select an option from possible Wikipedia pages

The annotation task requires two hours for our users to complete. Obviously, German annotators are more familiar with German culture than the Americans, and vice-versa. Annotators translate into their native language. Since we are focusing on popular entities, they are often known despite the cultural divide, but the introductory paragraph from Wikipedia reminds users if not.

6.5.2 Are the Adaptations Plausible?

To validate and compare all our adaptation strategies' precision, five German translators[5], and appropriately qualified **experts**, who understand American cul-

[5]Recruited through Upwork for $40 each.

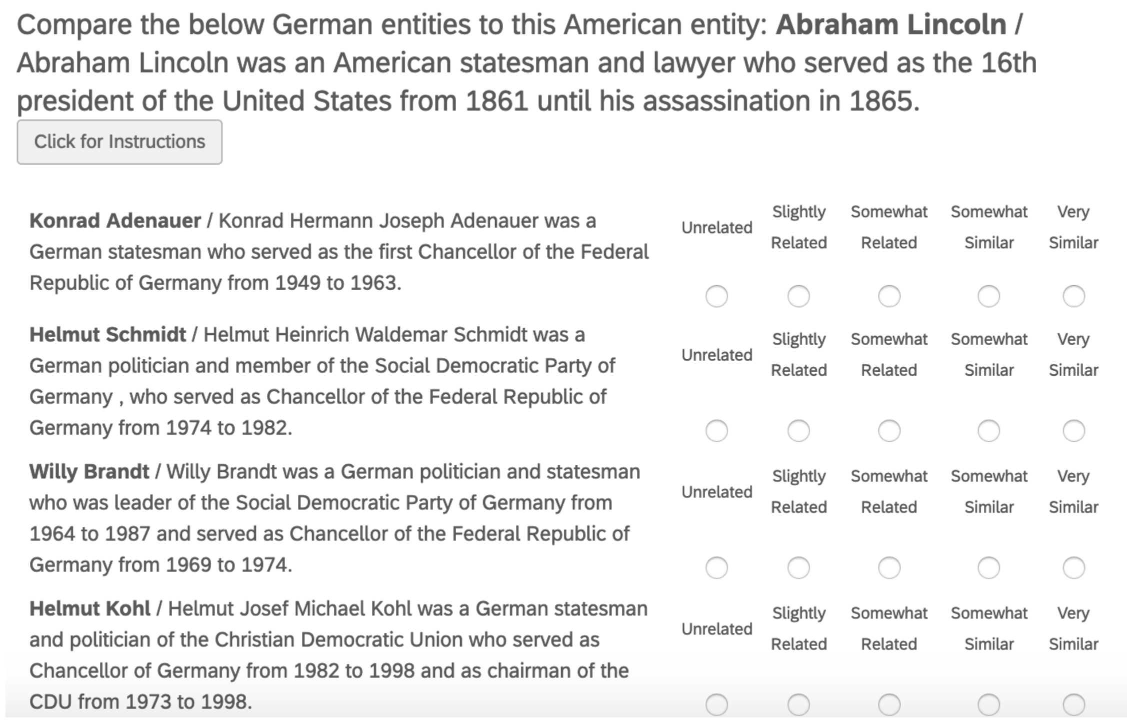

Figure 6.2: Our Qualtrics survey

ture assess the adaptations. The top five adaptations from WikiData, 3CosAdd, learned adaptation, and humans—as well as five randomly selected options from the human pool—are evaluated for plausibility on a five-level Likert scale.[6] Fleiss' Kappa (0.382) and Krippendorf's Alpha (0.381) assess interannotator Agreement; this "fair" agreement suggests that vetting an adaptation is challenging and sometimes subjective, even for translators.

[6] Our custom Qualtrics survey is provided in Figure 6.2. The order of adaptations is randomized and assessed on a Likert scale with anchors from Jurgens et al. (2014).

6.5.3 Why Adaptation is Difficult

Embedding adaptations are better than Wikidata's, and human adaptations are better still (Figure 6.3). Thus, we use human adaptations as the gold standard for evaluating recall. Only the learned embedding method uses training data, so we use human adaptations from Wikipedia to train the projection matrix and evaluate (for all methods) using human adaptations the NOC list. Given that the task is subjective, we take our results with a grain of salt given cultural variation (e.g., some people view Angela Merkel's conservatism as a defining characteristic, while others focus on her science pedigree).

We use the mean reciprocal rank (Voorhees et al., 1999, MRR) to measure how high the gold adaptations are ranked by our other adaptation strategies. Since MRR decreases geometrically and our gold standard is not exhaustive, the Recall@5, and @100 metrics are more intuitive. We calculate Recall@n by measuring what fraction of the correct adaptations of a source entity is retrieved in the top n predictions.[7] Table 6.2 validates that the human annotations are near the top of the automatic adaptations; the precision-oriented evaluation (Figure 6.3) validates whether the top of the list is reasonable. All human annotations and a sample of the automatic adaptations are provided in Appendix **??**.

[7] This is often referred to as P@n in bilingual lexicon induction literature (Conneau et al., 2017).

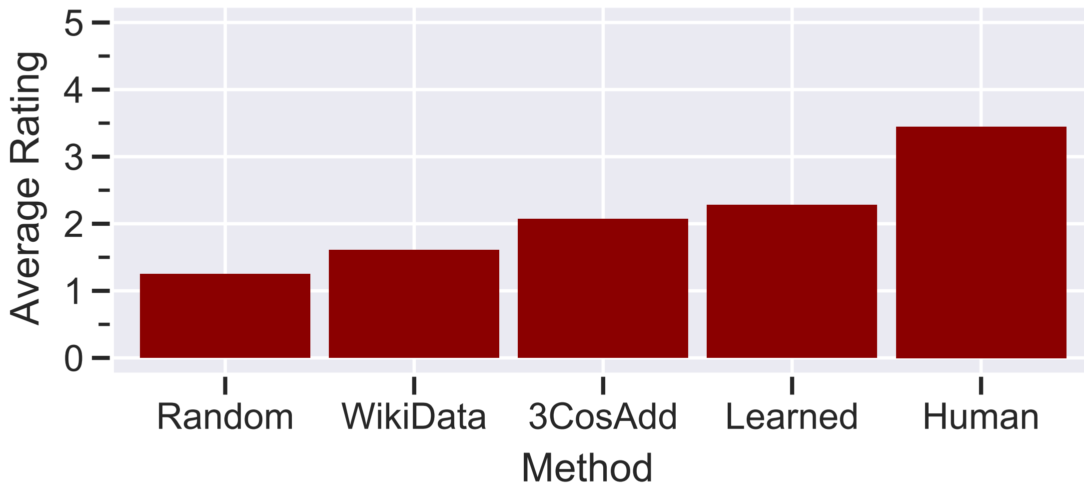

Figure 6.3: We validate adaptation strategies with expert translators on a five-point Likert scale. The human-generated adaptations are rated best—between "related" (3) and "similar" (4). These human adaptations become the reference for evaluation in Table 6.2.

6.5.4 Qualitative Analysis

There is no single answer to what makes a good adaptation. Let us return to the question of who Bill Gates is, which underlines how there is often no one right answer to this question but several context-specific possibilities. The human adaptations show the range of plausible adaptations, each appropriate for a particular facet of the position Bill Gates has in US society. As previously mentioned, Carl Benz represents a larger than life founder who created an entire industry with his company. However, Carl Benz made cars, not computers.

Even within technology, different adaptations highlight different aspects of Bill Gates. Like the implementer of the BASIC programming language, Konrad Zuse

Data	Metric	WikiData	3CosAdd	Learned
American→ *German*				
	Rec@5	7.5%	**14.2%**	-
Wikipedia	Rec@100	34.4%	**52.8%**	-
	MRR	0.05	**0.10**	-
	Rec@5	3.0%	22.9%	**28.6%**
Veale NOC	Rec@100	42.4%	**51.4%**	45.7%
	MRR	0.03	0.17	**0.24**
German→ *American*				
	Rec@5	3.1%	**17.2%**	-
Wikipedia	Rec@100	15.4%	**40.5%**	-
	MRR	0.01	**0.12**	-
	Rec@5	0.0%	**25.0%**	**25.0%**
Veale NOC	Rec@100	25.0%	**70.0%**	55.0%
	MRR	0.02	0.12	**0.15**

Table 6.2: If we consider human adaptations as correct, where do they land in the ranking of automatic adaptation candidates? In this recall-oriented approach, learned mappings (which use a small number of training pairs), rate highest.

contributed to computers that were more than single-purpose machines. Just as as Bill Gates's Microsoft is seen as a stodgy tech giant, Dietmar Hopp founded SAS, a giant German tech company that is more often discussed in board rooms than in living rooms. And because the epicenter of modern tech is America's West Coast, Andreas von Bechtolsheim represents a German founder of Sun Microsystems and early Google investor that made his way to Silicon Valley.

Other times, there is more consensus: a majority of raters declare Angela Merkel is the German Hilary Clinton, and Joseph Smith is the American Martin Luther. There are even some unanimous adaptations: Bavaria is the German California. Adaptations of fictional characters seem particularly difficult, although this may represent the supremacy of American popular culture; Superman and Homer Simpson are so well known in Germany that there are no clear adaptations; Till Eulenspiegel, Maverick, Bibi Blocksberg are not superheroes from a dying world and Heidi is not a dumb, bald everyman.

6.6 A New Computational Task

We formally introduce entity **adaptation** as a new computational task. Word2vec embeddings and WikiData can be used to figuratively—not just literally—translate entities into a different culture. Humans are better at generating candidates for this task than our computational methods (Figure 6.3). These methods are well-motivated, but have room for improvement. Knowledge bases improve over time and increased coverage of entities—as well as improved information about each entity—would improve the method. Alternate word embedding approaches—perhaps those that discard orthography—may provide better candidates. Even humans occasionally disagree with other humans on this task, so evaluation for this task is nontrivial.

Our new dataset of machine-generated adaptations, human adaptations, and human evaluation of these adaptations can serve as an evaluation for future automatic methods.

People need NLP systems that reflect their language **and** culture, but datasets are lacking: adaptation can help. There has been an explosion of English-language QA datasets, but other languages continue to lag behind. Several approaches try to transfer English's bounty to other languages (Lewis et al., 2019; Artetxe et al., 2019), but most of the entities asked about in major QA datasets are American (Gor et al., 2021). Adapting entire questions will require not just adapting entities and non-entities in tandem but will also require integration with machine translation (Kim et al., 2019; Hangya and Fraser, 2019). Our automatic methods did not create precise adaptations, but the alternative "incorrect" adaptations may be useful for low-precision tasks, such as generating numerous simple open-ended questions or gauging the popularity of an entity. Given the existence of robust datasets in high resource languages can we **adapt**, rather than literally translate, them to other cultures and languages?

This task is not possible without **expert annotation**. However, we do not **generate** full translations in this task. We do not observe malicious or careless answers from our **annotators** or **evaluators**. Hence, we extend the use of **experts** to a task in which quality assurance is nearly impossible: dialog **generation** in Chapter **??**.

Bibliography

Ravneet Arora, Chen-Tse Tsai, Ketevan Tsereteli, Prabhanjan Kambadur, and Yi Yang. 2019. A semi-Markov structured support vector machine model for high-precision named entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5862–5866, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted, connections to WordNet. In *Proc. Global WordNet Conference 2006*. Global WordNet Association.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Roy Freedle. 2003. Correcting the sat's ethnic and social-class bias: A method for reestimating sat scores. *Harvard Educational Review*, 73(1):1–43.

Hu Gengshen. 2003. Translation as adaptation and selection. *Perspectives: Studies in Translatology*, 11(4):283–291.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Maharshi Gor, Kellie Webster, and Jordan Boyd-Graber. 2021. Towards deconfounding the influence of subject's demographic characteristics in question answering. *arXiv preprint arXiv:2104.07571*.

Abigail Green. 2003. Representing germany? the zollverein at the world exhibitions, 1851–1862. *The Journal of Modern History*, 75(4):836–863.

Viktor Hangya and Alexander Fraser. 2019. Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234, Florence, Italy. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. 2019. Entity projection via machine translation for cross-lingual ner. *arXiv preprint arXiv:1909.05356*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. SemEval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 17–26, Dublin, Ireland. Association for Computational Linguistics.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. Low-resource deep entity resolution with transfer and active learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5851–5861, Florence, Italy. Association for Computational Linguistics.

David Katan and Mustapha Taibi. 2004. *Translating cultures: An introduction for translators, interpreters and mediators*. Routledge.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Maximilian Köper, Sabine Schulte im Walde, Max Kisselew, and Sebastian Padó. 2016. Improving zero-shot-learning for german particle verbs by using training-space restrictions and local scaling. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 91–96.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Tim Loh. Germany has its own dr. fauci—and actually follows his advice. *Bloomberg*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting Similarities among Languages for Machine Translation. *CoRR*, abs/1309.4.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (multidogo): Strategies toward curating and annotating large scale dialogue data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4518–4528.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 135–148.

Hagen Schulze. 1991. *The Course of German Nationalism: From Frederick the Great to Bismarck 1763–1867*. Cambridge University Press.

Peter D Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. *arXiv preprint arXiv:0809.0124*.

Tony Veale. 2016. Round up the usual suspects: Knowledge-based metaphor generation. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 34–41.

Jean-Paul Vinay and Jean Darbelnet. 1995. *Comparative stylistics of French and English: A methodology for translation*, volume 11. John Benjamins Publishing.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial question answering examples. *Transactions of the Association of Computational Linguistics*, 10.