

Corpus-Based Translation Induction in Indian Languages Using Auxiliary Language Corpora from Wikipedia

GOUTHAM THOLPADI, CHIRANJIB BHATTACHARYYA, and SHIRISH SHEVADE,
Indian Institute of Science

Identifying translations from comparable corpora is a well-known problem with several applications. Existing methods rely on linguistic tools or high-quality corpora. Absence of such resources, especially in Indian languages, makes this problem hard; for example, state-of-the-art techniques achieve a mean reciprocal rank of 0.66 for English-Italian, and a mere 0.187 for Telugu-Kannada. In this work, we address the problem of comparable corpora-based translation correspondence induction (CC-TCI) when the only resources available are small noisy comparable corpora extracted from Wikipedia. We observe that translations in the source and target languages have many topically related words in common in other “auxiliary” languages. To model this, we define the notion of a *translingual theme*, a set of topically related words from auxiliary language corpora, and present a probabilistic framework for CC-TCI. Extensive experiments on 35 comparable corpora showed dramatic improvements in performance. We extend these ideas to propose a method for measuring cross-lingual semantic relatedness (CLSR) between words. To stimulate further research in this area, we make publicly available two new high-quality human-annotated datasets for CLSR. Experiments on the CLSR datasets show more than 200% improvement in correlation on the CLSR task. We apply the method to the real-world problem of cross-lingual Wikipedia title suggestion and build the *WikiTSu* system. A user study on *WikiTSu* shows a 20% improvement in the quality of titles suggested.

CCS Concepts: • **Computing methodologies** → **Natural language processing**;

Additional Key Words and Phrases: Comparable corpora, translation correspondence induction, bilingual lexicon, cross-lingual semantic relatedness, auxiliary language, Wikipedia title suggestion

ACM Reference Format:

Goutham Tholpadi, Chiranjib Bhattacharyya, and Shirish Shevade. 2017. Corpus-based translation induction in Indian languages using auxiliary language corpora from wikipedia. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 16, 3, Article 20 (March 2017), 25 pages.
DOI: <http://dx.doi.org/10.1145/3038295>

1. INTRODUCTION

Identifying translations between two languages using comparable corpora is a well-known problem with several applications including cross-language retrieval and automatic or machine-assisted translation systems [Schafer and Yarowsky 2002]. This problem can be viewed as a special case of the broader problem of measuring cross-lingual semantic relatedness. In this work, we present an approach to translation induction from comparable corpora using auxiliary languages. The modeling assumptions made in our approach are seen to hold for the broader problem of semantic relatedness, and we present a method for computing cross-lingual semantic relatedness. These methods have wide applicability and we present their utility through a novel application:

Authors' addresses: G. Tholpadi, C. Bhattacharyya, and S. Shevade, Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560012, India; emails: gtholpadi@gmail.com, ([@csa.iisc.ernet.in](mailto:chiru.shirish)).

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2017 ACM 2375-4699/2017/03-ART20 \$15.00

DOI: <http://dx.doi.org/10.1145/3038295>

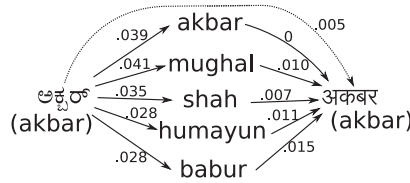


Fig. 1. A subset of the *translingual theme* in English (words in center) for a Kannada (left)–Marathi (right) translation pair. The arrow from w_1 to w_2 is labeled with the probability $P_{CC}(w_2|w_1)$ (see Section 3.2.1).

Wikipedia title suggestion. In what follows, we introduce the linguistic problems and the application problem and discuss the gaps in the current state of the art and the challenges involved.

1.1. Comparable Corpora-Based Translation Correspondence Induction

The task of identifying translations for terms is usually posed as one of generating translation correspondences. A translation correspondence for a source word assigns a score to every target word proportional to its topical similarity to the source word, so that the translation is assigned the highest score. Comparable corpora-based¹ translation correspondence induction (CC-TCI) is a popular approach for obtaining translation correspondences. Most methods use dictionaries and parsers or make assumptions about properties of the languages involved (see Section 2). However, for many language pairs such as in Indian languages, the CC-TCI problem poses several challenges:

- Resources such as seed bilingual lexicons and linguistic tools (POS taggers, morpho-syntactic analyzers, etc.) required by some methods (e.g., by Andrade et al. [2013] and Tamura et al. [2012], etc.) are not available.
- Language properties such as presence of cognates and orthographic similarity cannot be assumed in general, ruling out some methods (e.g., by Haghighi et al. [2008] and Koehn and Knight [2002], etc.).
- The only available cross-language resource is a comparable corpus. However, even this is relatively small for most language pairs, so that “CC-only” methods (e.g., by Ismail and Manandhar [2010] and Vulic et al. [2011], etc.) do not perform well.

We observe that source and target translations have many topically related words in common in other “auxiliary” language corpora,² which can be a useful cue for identifying translations. To model this, we define the notion of a *translingual theme* (for a source–target word pair) as a set of words derived from auxiliary language comparable corpora that statistically co-occur with the source and target words. For example, Figure 1 shows the source–target pair ಅಕ್ಬರ್/akbar/ and अकबर /akbar/ (both referring to the proper noun “Akbar”³) from a Kannada–Marathi corpus, and a subset {“mughal,” “shah,” “humayun,” “babur”}⁴ of its translingual theme derived from Kannada–English and Marathi–English auxiliary corpora. In this work, we investigate the utility of *auxiliary* language corpora for boosting CC-TCI performance. For this purpose, we leverage Wikipedia, a large web-based multilingual encyclopedia with more than 26 million articles in 285 languages. In Wikipedia, articles in different languages on the

¹In this article, the phrase “comparable corpora” is used to mean document-aligned multilingual corpora, where the aligned documents are written in different languages and “talk about the same thing” [Gaussier et al. 2004].

²Comparable corpora where one language is from the pair under consideration and the other can be any other (auxiliary) language.

³Akbar was a king from the Mughal dynasty who ruled parts of North India in the 16th century A.D.

⁴Shah is a royal title; Humayun and Babur were both Mughal kings.

same topic are linked (by “langlink”s), which enables us to quickly construct corpora for a large number of language pairs.

1.2. Cross-Lingual Semantic Relatedness

Translation induction can be viewed as a special case of the broader problem of measuring cross-lingual semantic relatedness (CLSR)—measuring the relatedness between words in different languages. Such measures can be applied to many cross-lingual tasks such as information retrieval (CLIR), text classification, machine translation, and lexicon induction [Hassan and Mihalcea 2009]. Measuring semantic relatedness between words in two languages is nontrivial. Previous approaches to this problem relied on the presence of rich Wikipedia “concept” inventories [Hassan and Mihalcea 2009] or rich linguistic tools like multilingual WordNets [Navigli and Ponzetto 2012] in the languages under consideration. However, such resources are not available in most languages, which makes the problem especially hard for resource-scarce languages, and very little work has been done in this area. Recently, Vulić and Moens [2013] proposed a “comparable corpora-only” approach to cross-lingual semantic relatedness, but they evaluated their method on the translation correspondence induction task due to the absence of evaluation datasets for semantic relatedness. In this work, we introduce two new human-annotated datasets for cross-lingual semantic relatedness. We also present an auxiliary language approach for measuring semantic relatedness. Experiments on the new datasets show that the proposed method improves semantic relatedness measurement. These datasets have been made publicly available to enable further research in this area.


1.3. Cross-Lingual Wikipedia Title Suggestion

To illustrate the utility of the approach on a real-world application, we consider the cross-lingual Wikipedia title suggestion problem, which requires solving the translation induction and the semantic relatedness problems effectively. This is a relatively new problem and especially relevant for multilingual societies. The problem stems from the fact that the proportion of content in Wikipedia in different languages varies widely [Udupa and Khapra 2010], and the topics covered also vary with language. If a Wikipedia concept has no article in one language, articles in other languages might be suggested to a multilingual user. For example (see Figure 2), an Indian user browsing the Kannada article ವೈರೂಷ /va:ʃra:ɳu/ (“virus”) might want to know about ಬ್ಯಾಕ್ಟೀರಿಯಾ /bja:kt̪i:rija/ (“bacteria”) and ರೋಟಾವೈರಸ್ /ro:taʋa:ʃras/ (“rotavirus”). There are no articles for these concepts in Kannada, but there *are* articles in Hindi, viz. जीवाणु /dʒi:ʋa:ɳu/ (“bacteria”) and रोटावाइरस /ro:ta:ʋa:i:ras/ (“rotavirus”). These titles can be suggested to the user (the box at top-right in the figure) for further reading. Recently, Bao et al. [2012] attempted a similar task using langlinks, where the setting was restricted to source words that are Wikipedia titles. This is because Wikipedia titles in one language have langlinks to titles in other languages, and this fact was leveraged in their approach. The task of suggesting target-language Wikipedia titles for *source words with no corresponding Wikipedia articles* is much harder and has not been attempted before. In the absence of langlinks, this task is difficult to solve, especially for underresourced languages without machine translation (MT), dictionaries, parsers, and parallel corpora. In this resource-scarce setting, we attempted the title⁵ suggestion task using a translation induction approach, leveraging auxiliary language document-aligned comparable corpora from Wikipedia. The resulting system *WikiTSu*

⁵In this article, we focus on Wikipedia titles that are single headwords. We plan to consider titles that are multiword units in future work.

ವೈರಾಣು

ವೈರಾಣುವು (ಬ್ಯಾಕ್ಟೀರಿಯಾ, ವೈರಸ್ ಎಂದೂ **ಟಾಕ್ಸಿನ್** - ಚೇದನವಿಲ್ಲದ ಉತ್ಪನ್ನದಾದ ವಿಷ ಅಥವಾ ವಿಷ) ಬೇರೆಯೊಂದು ಜೀವಿಯ ಜೀವಕೋಶಗಳೊಳಗೆ ಮಾತ್ರ ಸಂಕೀರ್ಣಗೊಳ್ಳುತ್ತಿ ಮಾಯವಲ್ಲ. ಒಂದು ಚಿಕ್ಕದಾದ ಸೂಕ್ಷ್ಮಜೀವಿಯಾದ **ಸೂಕ್ಷ್ಮಜೀವಿ**, ವೈರಾಣುಗಳು **ದ್ಯುಮೂಲಕ** ರೂಪದ ಉತ್ಪನ್ನದಾಗಿ ಕಾರಣವಾಗಿ ಚಿಕ್ಕದಾಗಿರುತ್ತವೆ. ವೈರಾಣುಗಳು **ಪ್ರೋಟೀನ್** ಮತ್ತು **ಸಸ್ಯಗಳ ಚೇದನಕಾರಿ (ಬ್ಯಾಕ್ಟೀರಿಯಾ)** ಮತ್ತು **ಆರ್ಕಿಯಾ** ದರ್ಗಾದ ವಿಷ ಬಗ್ಗೆಯ ಜೀವಾಣು, ಸೂಕ್ಷ್ಮಜೀವಿಗಳಾದುದು.



ರೋಟಾವೈರಸ್

ರೋಟಾವೈರಸ್ (ಬ್ಯಾಕ್ಟೀರಿಯಾ, ವೈರಸ್ ಎಂದೂ **ಟಾಕ್ಸಿನ್** - ಚೇದನವಿಲ್ಲದ ಉತ್ಪನ್ನದಾದ ವಿಷ ಅಥವಾ ವಿಷ) ಬೇರೆಯೊಂದು ಜೀವಿಯ ಜೀವಕೋಶಗಳೊಳಗೆ ಮಾತ್ರ ಸಂಕೀರ್ಣಗೊಳ್ಳುತ್ತಿ ಮಾಯವಲ್ಲ. ಒಂದು ಚಿಕ್ಕದಾದ ಸೂಕ್ಷ್ಮಜೀವಿಯಾದ **ಸೂಕ್ಷ್ಮಜೀವಿ**, ವೈರಾಣುಗಳು **ದ್ಯುಮೂಲಕ** ರೂಪದ ಉತ್ಪನ್ನದಾಗಿ ಕಾರಣವಾಗಿ ಚಿಕ್ಕದಾಗಿರುತ್ತವೆ. ವೈರಾಣುಗಳು **ಪ್ರೋಟೀನ್** ಮತ್ತು **ಸಸ್ಯಗಳ ಚೇದನಕಾರಿ (ಬ್ಯಾಕ್ಟೀರಿಯಾ)** ಮತ್ತು **ಆರ್ಕಿಯಾ** ದರ್ಗಾದ ವಿಷ ಬಗ್ಗೆಯ ಜೀವಾಣು, ಸೂಕ್ಷ್ಮಜೀವಿಗಳಾದುದು.

ರೋಟಾವೈರಸ್

<https://hi.wikipedia.org/s/9sot>
 ಮುಕ್ತ ಜ್ಞಾನಕೋಶ ವಿಕಿಪೀಡಿಯಾ ಸೇ

ರೋಟಾವೈರಸ್ ಛೋಟೆ ವರ್ಣಿಯೆಗೆ ಅತಿ ಸಾರ ಕಾರಣ ಕಾಲನ ಹೇ.^[1] ಇದು ಡಬಲ್-ಸ್ಟ್ರೆಂಡ್ಡ್ ಆರ್ ಎನ್ ಆ ವಿಶಾಲು ಕೀ ಆಕೃತಿ ಹೇ. ಲಗನು ಪಿಂಚು ವರ್ಷ ಕೀ ಆಯು ಮೆ ವಿಶು ಕೇ ಲಗನು ಸಮೀ ಬನ್ವೇ ರೋಟಾವೈರಸ್ ಸೇ ಕನು ಸೇ ಕನು ಆರ್ ಬಾರ ಅವಶ್ಯ ಸಂಕ್ರಮಿತ ಹೋತೆ ಹೇ.^[2]

ಗಲತೀ ಉದ್ಧತ ಕರ್ತೆ: <ref>ಟೇಗು ಮೊಜ್ದು ಹೇ, ಕಿನ್ಯು ಕೊಡೆ </references> ಟೇಗು ನಹೇ ಮಿಲ್ಲಾ

ಜೀವಾಣು

<https://hi.wikipedia.org/s/cwk>
 ಮುಕ್ತ ಜ್ಞಾನಕೋಶ ವಿಕಿಪೀಡಿಯಾ ಸೇ (ಬೆಂಟೋನಿ ಸೇ ಅನುಪೇಕ್ಷಿತ)

ಜೀವಾಣು ಆಕೃತಿ ಉದ್ದತೆ ಮಿಲಿಮೀಟರ್ ತಕ ಹೇ. ಇನಕೀ ಆಕೃತಿ ಗೋಲ ವಾ ಮುಕ್ತ-ಚಕ್ರಾಕಾರ ಸೇ ಲೇಕನು ಛಡಾ, ಆದಿ ಆಕಾರ ಕೀ ಹೋ ಸಕ್ತತೀ ಹೇ. ಏ ಪ್ರೋಕಾರಿಯೋಟಿಕ್, ಕೋಶಿಕಾ ಮಿಶ್ರಿತುಕ್, ಆರ್ಕಿಯೋಟಿಕ್ ಸರಲ ಜೀವ ಹೇ ಜೊ ಪ್ರಾಕ: ಸರ್ವತ್ರ ಪಾಯೆ ಜಾತೆ ಹೇ. ಏ ಪೃಥ್ವೀ ಪರ ಮಿಶ್ರಿ ಮೆ, ಅನ್ಯಲೀಯ ಗುಮೆ ಜಲ-ಧಾರಾಂ ಮೆ, ನಾಪಿಕೀಯ ಪದಾರ್ಥೊ ಮೆ^[1], ಜಲ ಮೆ, ಭೂ-ಪಪಡೀ ಮೆ, ಇಹಾಂ ತಕ ಕೀ ಕಾರ್ಬನಿಕ ಪದಾರ್ಥೊ ಮೆ ತಥಾ ಪಿಂಚೊ ಆಯು ಜನ್ಯುಂ ಕೇ ಶರೀರ ಕೆ ಖೀತರ ಮೊ ಪಾಯೆ ಜಾತೆ ಹೇ. ಸಾಧಾರಣತ: ಆಕ ಗ್ರಾಮ ಮಿಶ್ರಿ ಮೆ ೪ ಕನೇಡ ಜೀವಾಣು ಕೋಶ ತಥಾ ೧ ಮಿಲಿಲೀಟರ್ ಜಲ ಮೆ ೧೦ ಲಾಖ ಜೀವಾಣು ಪಾರೆ ಜಾತೆ ಹೇ. ಸಂಪೂರ್ಣ ಪೃಥ್ವೀ ಪರ ಅನುಮಾನತ: ಲಗನು 4x10³⁰ ಜೀವಾಣು ಪಾರೆ ಜಾತೆ ಹೇ.^[2] ಜೊ ಸಂಸಾರ ಕೇ ಬಾಯೆಮಾಸ ಕಾ ಆಕ ಬಹುತ ಬಡಾ ಪಾನ ಹೇ.^[3] ಏ ಕಾಡೆ ತಲ್ಯೊ ಕೇ ಘಕ್ರ ಮೆ ಬಹುತ ಮಹತ್ವಪೂರ್ಣ ಭೂಮಿಕಾ



Fig. 2. A multilingual user reading a Kannada article on ವೈರಾಣು (“virus”) (top-left) finds the words ಟಾಕ್ಸಿನ್ (“toxin”), ಬ್ಯಾಕ್ಟೀರಿಯಾ (“bacteria”), ಆರ್ಕಿಯಾ (“Archaea”) and ರೋಟಾವೈರಸ್ (“rotavirus”) interesting, but there are no Kannada articles for these concepts. In response, the system gives Wikipedia title suggestions (box at top-right) from Hindi and Tamil (ಜೀವಾಣು (“bacteria”), and so on).

can work for any Wikipedia language pair and uses a Wikipedia corpus as the *only* resource.

Contributions

Our main contributions are:

- We propose a comparable corpora-only approach for improving CC-TCI in under-resourced Indian languages using *translingual themes* derived from auxiliary language corpora. For this purpose, we define a new probabilistic notion of cross-language similarity in the context of comparable corpora. We show how this notion naturally admits auxiliary language corpora under certain assumptions. We also show how to combine similarities from multiple auxiliary languages using a simple mixture model and use the combined score for translation correspondence induction (Sections 3.2 and 3.3).
- We perform extensive experiments on 35 comparable corpora in nine languages from four language families (Indo-Aryan, Dravidian, Germanic, Romance) extracted from Wikipedia and show significant boosts (up to 124%) in performance for a state-of-the-art CC-TCI method (Section 6.1).
- We extend the auxiliary language approach to attack the problem of CLSR (Section 4). We show that the proposed method is significantly better at predicting semantic relatedness scores (up to 220% improvement in rank correlation) (Section 6.2).
- We introduce two new high-quality human-annotated datasets for evaluating CLSR (for Bengali-Marathi and Malayalam-Marathi; see Section 6.2.1 and Table XII).
- To address the cross-lingual Wikipedia title suggestion task for the difficult resource-scarce setting, we built a system *WikiTSu* that works for *any* language pair in

Table I. Acronyms and Notation

| Acronym/Symbol | Meaning |
|------------------------------|---|
| CC-TCI | Comparable corpora-based translation correspondence induction |
| CC-CLSR | Comparable corpora-based cross-lingual semantic relatedness |
| CC-only | Using only comparable corpora |
| TC | Translation correspondence |
| L_X | Language X |
| V_X | Vocabulary of language X |
| L_S, L_T, L_A | The source, target, and auxiliary languages |
| $S_{raw}^{TS}(\cdot, \cdot)$ | Scoring function that generates a TC between source and target languages |
| X | Random variable representing the word sampled from the L_X -document in a tuple of a document-aligned comparable corpus |
| x | The value taken by the random variable $X(x \in V_X)$ |
| $P_{CC}(\cdot, \cdot)$ | Trigger probability |
| $ST_A(s)$ | Source theme for the word $s \in V_S$ using the auxiliary language L_A |
| $TT_A(t)$ | Target theme for the word $t \in V_T$ using the auxiliary language L_A |
| $TLT_A(s, t)$ | Translingual theme for the order pair (s, t) |
| $P_A(\cdot, \cdot)$ | Auxiliary trigger probability |
| $S_A(\cdot, \cdot)$ | Final scoring function that incorporates auxiliary language information |
| AUX-COMB | The proposed method for CC-TCI that uses $S_A(\cdot, \cdot)$ |

Wikipedia, using *no other resources*. We show via a user study that *WikiTSu* does significantly better than a state-of-the-art baseline (Sections 5 and 6.3).

- We are releasing translation correspondences for 42 language pairs (nearly 5,000 words per language pair, 10 candidates per word) for public use as probabilistic dictionaries, as semantic networks, or as inputs to annotator tools for dictionary building. As of today, there exist *no* dictionaries for most of these language pairs (see supplementary material).
- We are making publicly available a large curated collection of comparable corpora and gold standard translation pair sets in seven underresourced languages. We are also releasing the code for *WiCCX*,⁶ a tool for generating preprocessed and algorithm-ready comparable corpora from Wikipedia dumps (see supplementary material).

2. RELATED WORK

2.1. Translation Correspondence Induction Using Comparable Corpora

The problem of inducing translation correspondences from bilingual comparable corpora was introduced by Rapp [1995]. There have been several approaches to this task, differentiated by the resource assumptions made.

Knowledge-Based Approaches. Many approaches to translation correspondence induction use seed lexicons, syntactic/morphological analyzers, parallel corpora, translation models, and other resources.

Andrade et al. [2013] use monolingual synonym sets to enrich the context vectors for words in different languages but depend on a bilingual dictionary with 1.6 million entries to make the context vectors comparable across languages. Prochasson and Fung [2011] and Irvine and Callison-Burch [2013] formulate the TCI problem as a classification problem that gives a binary decision for every source-target word pair and rely on seed lexicons in their approaches. On the other hand, our formulation

⁶Wikipedia Comparable Corpus Extractor.

induces a ranking over target words for a given source word, making it also useful for tasks that require the measure of relatedness between two words. They found that the most important feature for classification was the Wikipedia Topic feature that counts interlingually linked Wikipedia concepts where a word pair co-occurs. However, this dependency makes it difficult to apply on the relatively small and noisy Wikipedia corpora that are considered in this work. Klementiev et al. [2013] attack the larger problem of machine translation using bilingual lexicon induction as an integral step toward that goal. They use monolingual corpora but rely on a bitext (>2,500 sentence pairs) for tuning the model, and a bilingual dictionary (>49,000 entries). Laws et al. [2010] leverage linguistic similarities between translations and estimate similarity using a SimRank-based algorithm. The dependence on linguistic parsers (in each language) and a bilingual lexicon to connect the graphs in different languages make this method unsuitable for our setting. Qian et al. [2012] exploit similarities between translations in dependency relationships with other words. They, however, require dependency parsing tools in both languages and a seed lexicon. Laroché and Langlais [2010] discuss several methods for translation induction and find a common theme among them—constructing context vectors for each word, enriching the vectors using linguistic information, and using a seed lexicon to make the vectors comparable.

Other approaches make assumptions about the languages or corpora, such as syntactic structure, orthographic similarities, presence of cognates, monogenetic relationships, and domain-specific content [Rapp 1999; Laroché and Langlais 2010; Haghighi et al. 2008; Morin et al. 2008; Koehn and Knight 2002; Rubino and Linares 2011; Fišer and Ljubešić 2011]. Mausam et al. [2009] and Kaji et al. [2008] use existing dictionaries to induce translation correspondences. There is also work on comparable corpora-based named entity mining [Udupa et al. 2009; Li et al. 2011; Ji 2009], which has a similar setting but addresses a different problem. Udupa and Khapra [2010] use canonical correlation analysis for Wikipedia name search, and Erdmann et al. [2009] use Wikipedia link structure for translation correspondence induction, both of which are complementary to our statistical approach, and the methods can be combined to improve performance.

Comparable Corpora-Only Approaches. Ismail and Manandhar [2010] and Fung [1995] proposed methods that use only comparable corpora and were applied to relatively high-quality corpora. Rapp et al. [2012] use the WINTIAN algorithm [Rapp 1996] on document-level keywords extracted from aligned comparable corpora. The most recent work using only comparable corpora is by Vulić et al. [2011] and Vulić and Moens [2013], who use latent space models and demonstrate good performance on Wikipedia data.

Ravi and Knight [2011] introduced an interesting approach to learning translation tables from corpora that are not document aligned and later added optimizations to handle large-scale data [Ravi 2013]. This approach is useful when document-aligned corpora are not available. This work is complementary to our approach and the two can be combined to improve performance. For a good overview of current approaches to CC-TCI, the reader can refer to the survey by Sharoff et al. [2013].

Improving CC-TCI. There have been efforts to improve the results from existing methods by pre- or postprocessing. Li and Gaussier [2010] and Su and Babych [2012] attempt to improve corpus quality before doing translation correspondence induction. Shezaf and Rappoport [2010] take a noisy translation correspondence obtained from any method and incorporate knowledge from *monolingual corpora in the languages of the pair* to improve accuracy. Our method, on the other hand, takes a noisy translation correspondence and incorporates knowledge from *comparable corpora in auxiliary*

languages to improve accuracy. These approaches are complementary to our approach, and they can be combined to improve accuracy further.

2.2. Using Auxiliary Languages

Borin [2000] attempted to use auxiliary languages for translation correspondence induction, but using parallel corpora. Mann and Yarowsky [2001], Schafer and Yarowsky [2002], Mausam et al. [2009], and Tsunakawa et al. [2008] use existing dictionaries or monogenetic relationships, while we work in the comparable corpora-only setting and make no assumptions about the language family.

Auxiliary language approaches have also been used for other problems, for example, *triangulation* for machine translation [Wu and Wang 2007; Cohn and Lapata 2007; Utiyama and Isahara 2007; Dabre et al. 2014], word alignment [Kumar et al. 2007], transliteration [Khapra et al. 2010], paraphrase extraction [Bannard and Callison-Burch 2005], and so forth. Davidov and Rappoport [2009] use auxiliary languages for monolingual concept extension using bilingual lexicons. Banea et al. [2010] use auxiliary languages to improve subjectivity classification on English sentences but rely on machine translation. Hassan et al. [2012] use auxiliary languages to improve the estimation of semantic relatedness between texts in the same language.

While none of these methods are applicable to our setting due to the resource assumptions they make,⁷ we constructed two baselines based on the most recent work that was applicable, viz the methods proposed by Tsunakawa et al. [2008] (Section 6.1.4).

2.3. Cross-Lingual Semantic Relatedness Using Comparable Corpora

The problem of estimating CLSR was introduced by Hassan and Mihalcea [2009]. They leveraged the rich “concept” inventory of Wikipedia (each article is treated as a concept) and interlanguage links to derive a method similar to Explicit Semantic Analysis [Gabrilovich and Markovitch 2007]. However, this relies on a rich enough Wikipedia corpus, which is not available in most languages. For example, for the language pairs considered in our work, the size of the comparable corpora was usually less than 2,000 article pairs. Navigli and Ponzetto [2012] perform CLSR using a large multilingual semantic network incorporating knowledge from the English WordNet and Wikipedia—a resource unavailable to most resource-scarce languages. To the best of our knowledge, the only work on CLSR using comparable corpora only is by Vulić and Moens [2013]. They use a multilingual topic model to learn cross-lingual semantic similarity scores. Their setting is the closest to our work, and one of their methods (TI+Cue) is used as our baseline.

2.4. Combination Approaches

Laws et al. [2010] represent different kinds of relationships between words on a graph and use SimRank [Jeh and Widom 2002] to compute a combined score. Déjean et al. [2002] combine information with a mixture model similar to ours, while Rubino and Linares [2011] use a voting scheme instead.

3. APPROACH USING AUXILIARY LANGUAGES

3.1. Problem Formulation

For a given language L_X , let its vocabulary be denoted by the set V_X . Let L_S and L_T denote the source and target languages, with vocabularies V_S and V_T , respectively. The translation correspondence for $s \in V_S$ is the set $TC(s) = \{(t, r)\}_{t \in V_T}$, where $r \in \mathbb{R}$

⁷For example, Cohn and Lapata [2007] require parallel corpora to perform *triangulation*, and Mausam et al. [2009] require dictionaries in multiple language pairs for inducing translations for a single language pair.

is the topical similarity of t to s . A translation correspondence can be viewed as being generated from a scoring function $S_{raw}^{TS}(\cdot|\cdot) : V_T \times V_S \rightarrow \mathbb{R}_+$ such that $S_{raw}^{TS}(t|s) = r$. There exist several methods (Section 2.1) that can be used to learn a scoring function $S_{raw}^{TS}(t|s)$ from comparable corpora.⁸ This function induces a ranking over the words in V_T for each word s in V_S . We assume that there exists an auxiliary language L_A that has comparable corpora with L_S and L_T , so that we can learn scoring functions $S_{raw}^{AS}(a|s)$, $S_{raw}^{SA}(s|a)$, $S_{raw}^{TA}(t|a)$, and $S_{raw}^{AT}(a|t)$, analogous to $S_{raw}^{TS}(t|s)$.

The objective is to compute a scoring function $S_A(t|s)$ that incorporates the knowledge in the S_{raw}^{XY} scoring functions and gives a better ranking over V_T for each s .

3.2. Incorporating Information from an Auxiliary Language

3.2.1. Cross-Language Similarity in Terms of a Comparable Corpus. A document-aligned multilingual comparable corpus in l languages can be viewed as a set of tuples (each tuple contains l documents, one per language). Consider a random experiment where we sample a word from one of the documents of such a tuple. Define the following random variables: let S be the word sampled from the L_S -document in the tuple; let T be the word sampled from the L_T -document in the tuple. Let $P_{CC}(T = t|S = s)$ be the probability that the sampled L_T -word is t given that a sampled L_S -word is s . This probability will be high for some values of t (i.e., for some L_T -words) that are topically related to s . This follows from the following assumptions we have made:

- (1) All the documents in a tuple are on the same topic.
- (2) In a document on a particular topic, most of the words are related to that topic.⁹

Assumption 2 implies that s and t are likely to be topically related to the documents they are sampled from. Together with Assumption 1, this implies that s and t are likely to be topically related to each other. In other words, $P_{CC}(T = t, S = s)$ will be high for s and t that are topically related, and low for s and t that are unrelated. It follows that, for a given s , $P_{CC}(T = t|S = s) \propto P_{CC}(T = t, S = s)$ will be high when t is topically related to s and low for unrelated t .

For example, given that we sampled ಬ್ಯಾಕ್ಟೀರಿಯಾ /bja:ktiri:ja:/ (“bacteria”) from the L_S -document, we are very likely to sample words like जीवाणु /dʒi:vɑ:nu/ (“bacteria”) or रोग /ro:g/ (“disease”) from the L_T -document.¹⁰ We can use a baseline scoring function S_{raw}^{TS} and define the *trigger probability*.¹¹

$$P_{CC}(t|s) \triangleq \frac{S_{raw}^{TS}(t|s)}{\sum_{t'} S_{raw}^{TS}(t'|s)}.$$

The name “trigger” indicates the process where a target word t is triggered in response to a source word s given as a cue. This models topical relatedness in the context of comparable corpora.¹² Since this model is asymmetric (i.e. in general $P_{CC}(t|s) \neq P_{CC}(s|t)$), we can expect that the translation induction performance depends on the choice direction of induction, and this is confirmed by our experiments (Section 6.1.5).

3.2.2. Using Translingual Themes to Compute Word Similarity. The distributional hypothesis posits that “words that are similar in meaning occur in similar contexts” [Rubenstein

⁸We use the method by Vulić et al. [2011] to obtain S_{raw}^{TS} , and also as the baseline (Section 6.1.3).

⁹This assumption does not apply to function words. In our experiments, we remove function words from all documents as a preprocessing step.

¹⁰Here, $L_S =$ Kannada and $L_T =$ Hindi.

¹¹We abbreviate $P_{CC}(T = t|S = s)$ to $P_{CC}(t|s)$.

¹²This is different from $P_{MT}(t|s)$, the probability that a translator would consider that t is a translation of s , which is usually used in machine translation literature [Brown et al. 1993].

and Goodenough 1965]. This hypothesis is not directly applicable to words in different languages since they rarely occur together in a text. Many CC-TCI methods use a seed bilingual dictionary to map contexts (co-occurrence profiles) to a common space where they can be compared for words in different languages. Some methods use document alignment in comparable corpora to model document-level co-occurrence of words across languages. These methods rely on the quality and size of the corpora available. To augment the data available for such methods, we use comparable corpora in auxiliary languages. However, the new comparable corpus (in the auxiliary language) is not document-aligned with the original source-target comparable corpus, so that existing methods are not directly applicable. To handle this case, we build on the distributional hypothesis as follows.

Let us call words that occur in similar contexts as *distributionally similar* words. We hypothesize that *words that are semantically related are distributionally similar to the same set of words*; that is, if two words occur in similar contexts with the same set of words, then the two words are topically related. Given a source language word s , we can get distributionally similar words in an auxiliary language using the document-aligned comparable corpus¹³ in the source and auxiliary languages. We call this the *source theme*. Similarly, we get auxiliary language words that are distributionally similar to a target language word t —the *target theme*. If the source and target themes have significant overlap, then it boosts our confidence that s and t are semantically related.

Consider the example in Figure 1 of a Kannada–Marathi translation pair ಅಕ್ಕಬರ–अकबर meaning the proper noun “Akbar.” Using the Kannada–Marathi Wikipedia corpus, we get a low value of $P_{CC}(\text{अकबर}|\text{ಅಕ್ಕಬರ}) = .005$. Using document-aligned comparable corpora in Kannada–English and Marathi–English, we obtain distributionally similar English words for the source and the target words to get the source and target themes, respectively. The intersection of the two yields the translingual theme consisting of auxiliary language words a (e.g., “akbar,” “mughal,” “shah,” “humayun,” and “babur”), which have a high value for both $P_{CC}(\text{अकबर}|a)$ and $P_{CC}(a|\text{ಅಕ್ಕಬರ})$. This evidence can be used to boost the score for the target candidate अकबर.

We formalize this intuition in a probabilistic framework as follows. We start with the setting of a tuple of aligned comparable documents described in Section 3.2.1. (We will relax this requirement later in this section.) Let A be a random variable representing the word sampled from the L_A -document in the tuple. Then, similar to $P_{CC}(t|s)$, we can get $P_{CC}(t|a)$ and $P_{CC}(a|s)$, $\forall a \in V_A$. Our probabilistic definition allows us to write $P_{CC}(t|s) = \sum_{a \in V_A} P_{CC}(t|a, s)P_{CC}(a|s)$. This formulation allows us to use information from auxiliary languages as follows. By assuming that T is independent of S given A , we can define the *auxiliary trigger probability*,^{14,15}

$$P_A(t|s) \triangleq \sum_{a \in V_A} P_{CC}(t|a)P_{CC}(a|s). \quad (1)$$

The independence assumption means that we are no longer constrained to use a multilingual corpus but can use several bilingual corpora—one for each language pair. This addresses the problem of lack of document alignment between the new auxiliary

¹³This corpus need not be document aligned with the original corpus.

¹⁴While this equation looks identical to the triangulation equation described by Cohn and Lapata [2007], the underlying probabilistic model there is $P_{MT}()$ (see footnote 9), while in our case it is $P_{CC}()$.

¹⁵If a word a is not present in the L_A – L_T corpus, we need to use a noninformative uniform back-off distribution for $P_{CC}(t|a)$ (as suggested by Cohn and Lapata [2007] for dissimilar corpora).

language corpus and the original corpus. This is critical, since multilingual corpora are far more difficult to obtain than bilingual corpora.¹⁶

Translingual themes. Summing over the entire auxiliary language vocabulary V_A in Equation (1) introduces a lot of noise [Ismail and Manandhar 2010] and is computationally expensive. We need a more focused and reliable indicator of topical relatedness. For this purpose, we construct a *translingual theme* for a given word pair—a set of words in the auxiliary language that are highly related to both the source and the target words—and use that in the previous formulation.

We define the *source theme* for s as

$$ST_A(s) = \left\{ a \in V_A : \begin{array}{l} \text{(i) } \forall a' \in V_A - ST_A(s), \text{ we have } P_{CC}(a|s) \geq P_{CC}(a'|s). \\ \text{(ii) } \sum_{a \in ST_A(s)} P_{CC}(a|s) \geq \tau, \text{ but for any } b \in ST_A(s) \\ \text{we have } \sum_{a \in ST_A(s) \setminus b} P_{CC}(a|s) < \tau. \end{array} \right\}. \quad (2)$$

In other words, we arrange the auxiliary language words in the descending order of trigger probability, start selecting the top-ranked words, and stop when the total probability mass of the selected words reaches a threshold τ . The ordering step ensures that we give preference to highly related words, and the thresholding step prevents the inclusion of noise by excluding low-probability words. We use this kind of cumulative probability thresholding instead of simply choosing the top k words for two reasons: (1) if a few auxiliary language words are highly related (and carry most of the probability mass), then the top k would contain low-probability noise words, and (2) if we have a relatively large number of related auxiliary language words (each with moderate probability mass), then we would lose this information if we use only the top k words. Using cumulative probability thresholding mitigates both of these problems.

The threshold τ takes a value in the range $[0, 1]$. A value closer to 0 allows a few highly related words into the source theme but makes it difficult to compare with the target theme if there is not enough overlap. A value closer to 1 allows a large chunk of the auxiliary language vocabulary into the source theme, increasing noise, and also makes the computation very slow. The best value for the threshold $\tau \in [0, 1]$ can be determined empirically.¹⁷

A symmetric definition for the *target theme* for t can be obtained by replacing $P_{CC}(a|s)$ by either $P_{CC}(a|t)$ or $P_{CC}(t|a)$ in Equation (2). Since Equation (1) uses $P_{CC}(t|a)$, and because P_{CC} is not symmetric (as discussed in Section 3.2.1), we should not use $P_{CC}(a|t)$ in the definition of the target theme. Using $P_{CC}(t|a)$ is also problematic because, for two different auxiliary language words a and a' , the probabilities $P_{CC}(t|a)$ and $P_{CC}(t|a')$ are derived from two different distributions (determined by the conditioning) and cannot really be compared or summed up, as required by the definition in Equation (2). For these reasons, we define the target theme for t as

$$TT_A(t) = \{a \in V_A : t \in ST_T(a)\};$$

that is, we take auxiliary language words that have t in their *source themes*. In other words, these are auxiliary language words for which t has a high trigger probability. Finally, we define the *translingual theme* for the ordered pair (s, t) as

$$TLT_A(s, t) = ST_A(s) \cap TT_A(t).$$

For the example in Figure 1, $TLT_{\{en\}}(\text{ಅಕಬರ್, ಮುಘಲ್, ಶಾಹ}) = \{\text{“akbar,” “mughal,” “shah,” “humayun,” “babur”}\}$. The auxiliary trigger probability given in Equation (1) is now

¹⁶As shown in the supplementary material, the trilingual corpora formed using the language pair and the auxiliary language were too small to learn useful topic models.

¹⁷We found that a value of τ in the range $[0.7, 0.9]$ gave good performance.

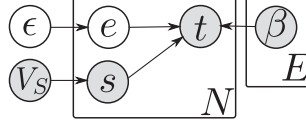


Fig. 3. Simple combining model.

redefined as

$$P_A(t|s) \triangleq \sum_{a \in TLT_{A(s,t)}} P_{CC}(t|a)P_{CC}(a|s). \quad (3)$$

In the rest of this article, we will use the previous definition of auxiliary trigger probability. We use $P_A(t|s)$ as a measure of the topical similarity between t and s . For the example in Figure 1, we compute $P_A()$ as follows:

$$\begin{aligned} & P_{\{\text{en}\}}(\text{अकबर}|\text{अकबर}) \\ &= \sum_{a \in TLT_{\{\text{en}\}}(\text{अकबर})} P_{CC}(\text{अकबर}|a)P_{CC}(a|\text{अकबर}) \\ &= P_{CC}(\text{अकबर}|\text{mughal})P_{CC}(\text{mughal}|\text{अकबर}) + \dots + P_{CC}(\text{अकबर}|\text{babur})P_{CC}(\text{babur}|\text{अकबर}) \\ &= .041 \times .010 + \dots + .028 \times .015 = .0014. \end{aligned}$$

3.3. Model for Combining Languages

Since both $P_{CC}(t|s)$ and $P_A(t|s)$ are imperfect indicators of translation correspondence, we would like to combine both scores but weight the contribution of each distribution according to its performance on a small development set. Consequently, we chose a simple mixture model for combining information. The generative story for the model (Figure 3) is as follows:

- (1) Sample a source word s uniformly from the source vocabulary V_S .
- (2) For each s :
 - (a) Sample $e \sim \text{Discrete}(\lambda)$. (e is one of the mixture components.)
 - (b) Sample $t \sim \text{Discrete}(\beta_{es})$. (A mixture component is a discrete distribution over the target vocabulary.)

We can learn the distributions $P_0(t|s) \triangleq P_{CC}(t|s)$ and $P_j(t|s) \triangleq P'_{A_j}(t|s)$, $j = 1 \dots E$ for the auxiliary language set $A = \{A_j\}_{j=1}^E$ using a set of comparable corpora.¹⁸ We have

$$\begin{aligned} p(t|s, \lambda) &= \sum_e p(t, e|s, \lambda) = \sum_e p(e|s, \lambda)p(t|e, s, \lambda) = \sum_{j=1}^E p(e = j|\lambda)p(t|e = j, s, \lambda) \\ &= \sum_{j=0}^E \lambda_j \beta_{jst}, \end{aligned}$$

where $\beta_{jst} \triangleq P_j(t|s)$ and $\lambda_j \triangleq p(e = j|\lambda) \geq 0$ for $j = 0 \dots E$, and $\sum_{j=0}^E \lambda_j = 1$. Given a small development set of source-target translation pairs $\{(s_n, t_n)\}_{n=1}^N$,¹⁹ we can learn λ by grid search or by maximizing the log-likelihood $\sum_n \log \sum_j \lambda_j \beta_{j s_n t_n}$ w.r.t. λ , subject to

¹⁸In our experiments, we have tried $E = 1, 2$, and 3 .

¹⁹Note that this development set of a few (<100) translation pairs is different from the seed lexicons mentioned in Section 1, which are bilingual lexicons of a few thousand translation pairs that are used by some methods (e.g., by Tamura et al. [2012]) to bootstrap cross-language comparisons. We do not use such seed lexicons.

the constraints mentioned previously. For the maximum likelihood approach, we used the Expectation–Maximization (EM) algorithm.²⁰ We initialize λ randomly, and then use the following updates till convergence:

$$b_{nj} = \frac{\beta_{js_n t_n} \lambda_j}{\sum_{j'} \beta_{j' s_n t_n} \lambda_{j'}}, \quad \lambda_j = \frac{\sum_n b_{nj}}{\sum_{j'} \sum_n b_{nj'}}.$$

We do multiple random initializations and keep the λ with the best likelihood. Having learned λ , we can compute $p(t|s, \lambda)$ for any word pair (s, t) . The **new scoring function** $S_A(\cdot)$ is defined as $S_A(t|s) \triangleq p(t|s, \lambda) = \sum_{j=1}^E \beta_{jst} \lambda_j$. The translation correspondence for s is defined as $TC(s) = \{(t, S_A(t|s))\}_{t \in V_T}$, and the translation candidate t^* for s is defined as $t^* = \arg \max_t S_A(t|s)$. Through β , other cues can also be introduced, for example, other scoring functions on the same corpus, limited-coverage dictionaries, and multilingual WordNets.

In the example in Figure 1, $P_0(\text{अकबर}|\text{अक़्ब_0}) = P_{CC}(\text{अकबर}|\text{अक़्ब_0}) = .005$. This is the score from the baseline method TI+Cue (or S_{raw}^{TS}). $P_1(\text{अकबर}|\text{अक़्ब_0}) = P_{\{\text{en}\}}(\text{अकबर}|\text{अक़्ब_0}) = .0014$. The estimated value of λ from the training pairs was $(\lambda_0 = 0.37, \lambda_1 = 0.63)$. The score from the proposed scoring function is

$$\begin{aligned} S_{\{\text{en}\}}(\text{अकबर}|\text{अक़्ब_0}) &= \lambda_0 P_0(\text{अकबर}|\text{अक़्ब_0}) + \lambda_1 P_1(\text{अकबर}|\text{अक़्ब_0}) \\ &= .37 \times .005 + .63 \times .0014 = .0027. \end{aligned}$$

Note that even though $S_{\{\text{en}\}}(\text{अकबर}|\text{अक़्ब_0}) = .0027 < .005 = P_{CC}(\text{अकबर}|\text{अक़्ब_0})$, the rank of the translation अकबर changes from 6 (using P_{CC}) to 3 (using $S_{\{\text{en}\}}$). This is because using the auxiliary language information helps us to reject false positives among the translation candidates generated by P_{CC} by shrinking their $S_{\{\text{en}\}}$ score much more than for the translation.

4. MODELING SEMANTIC RELATEDNESS

The trigger probability $P_{CC}(t|s)$ models the topical relatedness between words in different languages, as described in Section 3.2, and is not just a probability of translation. This probability is conceptually similar to the conditional probability defined for cross-lingual *lexical triggers* described by Kim and Khudanpur [2004], where the authors model the fact that the presence of a particular word in a document signals the existence of topically related words in a comparable document in another language. This idea also relates to the definition of semantic relatedness in terms of *semantic word responses* proposed by Vulić and Moens [2013]. The authors describe the process of generating semantically related words in terms of the process by which humans produce words as free word associations given some cue word. This suggests that the trigger probability $P_{CC}(\cdot|\cdot)$ could be used for modeling semantic relatedness as well.

Kim and Khudanpur [2004] argue that the trigger probability is closely related to statistical measures such as mutual information. This is consistent with previous work that successfully used pointwise mutual information (PMI) as a measure of semantic relatedness [Turney 2001; Bollegala et al. 2007]. In our setting, PMI is not directly applicable as the computation of the terms $p(s, t)$, $p(s)$, and $p(t)$ as described in previous work is not possible. Hence, we used the available information (trigger probabilities) to arrive at a formulation that is conceptually similar to PMI. For a given word pair (s, t) in source and target languages, we can get the trigger probabilities $p(t|s)$ and $p(s|t)$. The semantic relatedness score SR for (s, t) is defined as $SR(s, t) \triangleq \log p(t|s)p(s|t)$. This

²⁰We report results using the grid search in the article and the results using the EM algorithm in the supplementary material.

can be viewed as a function of the PMI as follows:

$$SR(s, t) = \log p(t|s)p(s|t) = \log \frac{p(t, s)}{p(s)} \frac{p(s, t)}{p(t)} = \log p(s, t) + PMI(s, t).$$

This formulation also ensures that:

- (1) the semantic relatedness score is symmetric (i.e., $SR(s, t) = SR(t, s)$), and
- (2) we do not need to compute any quantities that are not part of our model (e.g., $p(s, t)$, $p(s)$, $p(t)$, etc.). The trigger probabilities are the only quantities required.

5. APPLICATION TO CROSS-LINGUAL WIKIPEDIA TITLE SUGGESTION

In cross-lingual Wikipedia Title suggestion, the input is a word in the source language and the task is to generate a word in the target language that is also a Wikipedia title. Since the target translation may not always be the title of a Wikipedia article, the requirement is to generate a word that is most related to the source word. Thus, the task has elements of both the translation correspondence induction task (generation of a single response word in the target language) and the cross-lingual semantic relatedness task (generation of semantically related words in other languages). This task is similar in spirit to the *semantic word responding* activity described in Section 4. This suggests the use of *trigger* probabilities to generate the title suggestion. The semantic word responding activity and the title suggestion task are not symmetric. For example, the first response to “Neil Armstrong” might be “moon,” but the first response to “moon” is unlikely to be “Neil Armstrong.” This is naturally modeled using the trigger probabilities, and the symmetrization achieved by the semantic relatedness score $SR(s, t)$ is not useful here. Hence, we used the following procedure to generate the Wikipedia title suggestion: given the source word s , sort the target words t in descending order of trigger probability $p(t|s)$ and choose the first word t that is also a Wikipedia title. In the current work, we have focused on Wikipedia titles that are single headwords. We plan to consider titles that are multiword units in future work.

6. EXPERIMENTS AND RESULTS

We present three sets of experiments to address the three main tasks discussed in the preceding sections. In Section 6.1, we discuss the results on comparable corpora-based translation correspondence induction. The datasets described in this section were also used for the other tasks. The newly created datasets for cross-lingual semantic relatedness and the correlation experiments are described in Section 6.2. Finally, we describe a user study to evaluate the performance of the *WikiTSu* system on the cross-lingual Wikipedia title suggestion task in Section 6.3. In the remainder of this section, we refer to our method for CC-TCI as **AUX-COMB**, our method for CC-CLSR as **AUX-COMB-SR**, and our method for Wikipedia title suggestion as **AUX-COMB-WTS**. Analogously, we refer to the application of the TI+Cue method to the three tasks as TI+Cue, TI+Cue-SR, and TI+Cue-WTS, respectively.

6.1. Translation Correspondence Induction

We evaluated the AUX-COMB method on 21 language pairs derived from seven Indian languages from two language families—Indo-Aryan: Bengali (bn), Hindi (hi), and Marathi (mr); and Dravidian: Kannada (kn), Malayalam (ml), Tamil (ta), and Telugu (te). We used three auxiliary languages from different language families—Germanic: English (en), Romance: French (fr), and Indo-Aryan: Hindi. We extracted 35 comparable corpora (624,856 documents in total) from Wikipedia, which were the largest corpora possible (using all available `langlinks`). We used a state-of-the-art method for TCI to measure the impact of incorporating auxiliary languages.

6.1.1. Corpus and Gold Standard. We used articles from Wikipedia in nine languages²¹ as the corpus for our experiments. The data was processed using *WiCCX*, a tool for preparing Wikipedia corpora. The tool was also used to compile translation pairs using *langlinks* to use as the gold standard for evaluating CC-TCI. More details about the corpora and the gold sets, the steps involved in their preparation, and some analysis about the size and quality of the data are given in the supplementary material.

6.1.2. Evaluation Procedure. The gold-standard translation pair sets for some of the language pairs were quite small. To mitigate this problem, we used Monte Carlo cross-validation, which has been shown to be asymptotically consistent [Picard and Cook 1984], resulting in more pessimistic predictions of the test data compared with normal cross-validation. The gold sets were divided into development and test sets in k different ways by random sampling.²² The size of the development set d was fixed at 40²³ for all language pairs, and the remaining translation pairs were used for testing.

Evaluation Measures. Given a test set in languages L_1 and L_2 , for each word in L_1 in the test set, each method was used to generate a ranked list of candidate words in L_2 . Similarly, L_1 candidates were generated for L_2 words. Each ranked list was evaluated in terms of mean reciprocal rank (MRR) [Voorhees 1999].²⁴ Let $tr(w)$ be the translation of w in the gold set. Given a ranked list generated for w , $RR(w) = \frac{1}{\text{Rank of } tr(w) \text{ in the list}}$. The reciprocal ranks were averaged over all words in the test set and then averaged over all k folds in the cross-validation to get the final score.

Since the gold sets differed between experiments, the scores are not directly comparable. Instead, we report performance improvement over the baseline score (computed on the same gold set).²⁵

6.1.3. Scoring Function and Baseline. Given the noisy nature of the Wikipedia corpus, we chose the **TI+Cue** method as our baseline. The TI+Cue method is a state-of-the-art method for CC-TCI, proposed by Vulić et al. [2011]. It is based on topic models [Mimno et al. 2009], which work at the coarser level of topics (rather than words, or documents), and hence can be expected to smooth out noise better. This method also yielded the scoring functions S_{raw}^{XY} (as described in Section 3.1), which was used by AUX-COMB and is described in the supplementary material.

For bilingual topic modeling, we used the Mallet toolbox²⁶ with the following configuration: regex for importing data = “[\p{L}\p{M}]+”, number of topics $K = \lceil \frac{\#doc \text{ pairs}}{10} \rceil$, $\alpha = \frac{50}{K}$, $\beta = 0.01$, number of iterations = 1,000 (estimation) and 100 (inference), and burn-in period = 100 iterations.

6.1.4. Auxiliary Language Method Baselines. As discussed in Section 2.2, previous approaches to using auxiliary languages used resources that are not available in our setting. The available resources in our setting are (1) comparable corpora and (2) translation correspondences induced from the corpora by a state-of-the-art CC-TCI method (TI+Cue). Using these resources, we constructed two baselines as follows:

²¹<http://dumps.wikimedia.org/>.

²²We fixed $k = 10$ in our experiments.

²³We set $d = 35$ for $A = \{en, fr\}$, and $d = 25$ for $A = \{en, fr, hi\}$, since some language pairs had very small reduced gold sets.

²⁴We also measured “Presence-at- k ” (Pres@ k) for $k = 1$ and 5. In general, these measures showed the same trends as MRR. The details are given in the supplementary material.

²⁵We report the absolute scores for the baseline on $G(\{en\})$ in Table II to give the reader an idea of the absolute MRR scores. The absolute scores for all cases are reported in the supplementary material.

²⁶<http://mallet.cs.umass.edu>.

Table II. Absolute Performance (in Terms of MRR) of TI+Cue on the English Gold Set $G(\{en\})$ (Poorly Performing Language Pairs Are in Bold)

| MRR | bn | hi | kn | ml | mr | ta | te |
|-----|-------|-------|--------------|--------------|-------------|-------|-------|
| bn | – | .3174 | .1842 | .2422 | .2439 | .2923 | .2271 |
| hi | .284 | – | .2837 | .2408 | .3145 | .283 | .2942 |
| kn | .2113 | .2966 | – | .1273 | .165 | .2342 | .2313 |
| ml | .2500 | .3228 | .1522 | – | .2226 | .2416 | .2381 |
| mr | .2230 | .349 | .1403 | .1876 | – | .2832 | .2488 |
| ta | .2731 | .3232 | .241 | .2472 | .2511 | – | .2483 |
| te | .2506 | .2943 | .1748 | .3543 | .2318 | .2571 | – |

ALB-SS. Given the source-auxiliary translation correspondence $TC_A(s) = \{(a, P_{CC}(a|s))\}_{a \in V_A}$, we take $a^* = \arg \max_{a \in V_A} P_{CC}(a|s)$ and set the source-target translation correspondence $TC_T(s) = TC_T(a^*) = \{(t, P_{CC}(t|a^*))\}_{t \in V_T}$; that is, we take the top-ranked auxiliary language translation and use its target translation correspondence. We call this the auxiliary language baseline where we have a *strong source* (ALB-SS); that is, the source-auxiliary translation correspondence is treated as strong evidence.

ALB-ST. Given the auxiliary-target translation correspondence $TC_T(a) = \{(t, P_{CC}(t|a))\}_{t \in V_T}$, we take $t_a^* = \arg \max_{t \in V_T} P_{CC}(t|a)$ for each $a \in V_A$ and set the source-target translation correspondence $TC_T(s) = \{(t_a^*, P_{CC}(a|s))\}_{a \in V_A}$. In other words, we use the source-auxiliary translation correspondence but replace each auxiliary language word by its top-ranked target language translation. We call this the auxiliary language baseline where we have a *strong target* (ALB-ST); that is, the auxiliary-target translation correspondence is treated as strong evidence.

It is natural to conceive of a baseline where both the source and target evidence are treated as strong evidence. This treatment effectively results in bilingual dictionaries induced by taking the top-ranked candidate from every translation correspondence. More precisely, we construct a source-auxiliary bilingual lexicon $\{(s, a)\}_{s \in V_S, a \in V_A}$ by collecting all word pairs (s^*, a^*) that satisfy either $a^* = \arg \max_{a \in V_A} P_{CC}(a|s^*)$ or $s^* = \arg \max_{s \in V_S} P_{CC}(s|a^*)$. Similarly, we construct a target-auxiliary bilingual lexicon. Given this additional resource, we can consider dictionary-based auxiliary language methods for comparison. The work by Tsunakawa et al. [2008] is the most recent work in this area.²⁷ We consider two of the three methods proposed by them as baselines, viz. *exact merging* (**TsuEM**) and *alignment-based merging* (**TsuAM**).²⁸ The reader is referred to the paper by Tsunakawa et al. [2008] for the details of these two methods.²⁹

6.1.5. Discussion of Results. The performance of the TI+Cue method for $G(en)$ is shown in Table II (also see Section 6.1.2). The number in row L_S and column L_T is the performance measured when identifying translations for L_S words in L_T . It can be seen that MRR is in the range [0.2,0.3] for most language pairs, and even lower for *bn-kn*, *kn-ml*, *kn-mr*, and *ml-mr*, which have small corpora sizes (<1,000). We believe that using auxiliary language corpora will be especially useful for such language pairs.

²⁷The methods proposed by Mausam et al. [2009] mainly leverage the existence of multiple dictionaries in several languages in order to induce translations for a particular language pair. This is not available in our setting.

²⁸The third method, viz. *word-based merging*, is identical to exact merging in our case, since we consider only single-word units in this article.

²⁹In the case of alignment-based merging, we used a large monolingual Wikipedia corpus to estimate the monolingual language model, rather than using the Google hit count as suggested by the authors.

Table III. Absolute Performance (in Terms of MRR) of ALB-SS on $G(en)$

| MRR | bn | hi | kn | ml | mr | ta | te |
|-----|-------|-------|-------|--------|-------|-------|-------|
| bn | – | .1209 | .1258 | 0.0699 | .0831 | .0326 | .099 |
| hi | .0987 | – | .0629 | .0936 | .0677 | .0769 | .1009 |
| kn | .0792 | .0805 | – | .0386 | .0232 | .0442 | .1018 |
| ml | .0878 | .1136 | .0513 | – | .0538 | .0494 | .0765 |
| mr | .0854 | .1047 | .0489 | .0685 | – | .0208 | .0545 |
| ta | .0786 | .0598 | .051 | .0608 | .0199 | – | .058 |
| te | .0744 | .0813 | .0894 | .0915 | .035 | .04 | – |

Table IV. Absolute Performance (in Terms of MRR) of ALB-ST on $G(en)$

| MRR | bn | hi | kn | ml | mr | ta | te |
|-----|-------|-------|-------|-------|-------|-------|-------|
| bn | – | .0425 | .0687 | .0307 | .0467 | .0234 | .0486 |
| hi | .0224 | – | .0378 | .0228 | .0492 | .0461 | .0353 |
| kn | .0432 | .029 | – | .0146 | .0602 | .0352 | .0559 |
| ml | .0349 | .0223 | .0098 | – | .0329 | .0079 | .0072 |
| mr | .0616 | .0781 | .0432 | .0184 | – | .0148 | .0154 |
| ta | .0379 | .033 | .0191 | .0093 | .0217 | – | .0274 |
| te | .0456 | .0452 | .0343 | .0329 | .0485 | .0417 | – |

Table V. Absolute Performance (in Terms of MRR) of TsuEM on $G(en)$

| MRR | bn | hi | kn | ml | mr | ta | te |
|-----|-------|-------|-------|-------|-------|-------|-------|
| bn | – | .0446 | .1017 | .0223 | .0378 | .0384 | .0528 |
| hi | .0354 | – | .0419 | .0342 | .0491 | .0511 | .0332 |
| kn | .1774 | .0759 | – | .025 | .0463 | .038 | .0705 |
| ml | .056 | .0733 | .0086 | – | .0323 | .0574 | .0181 |
| mr | .0466 | .0801 | .0088 | .0311 | – | .0316 | .0334 |
| ta | .0242 | .0245 | .0304 | .0154 | .0299 | – | .0499 |
| te | .1021 | .0489 | .0434 | .0244 | .0149 | .1163 | – |

Auxiliary Languages Boost Performance. The performance of the baseline auxiliary language methods for $G(en)$ is given in Tables III, IV, V, and VI. We see poor performance when compared to TI+Cue. We feel this is because choosing only the top-ranked word from the auxiliary language vocabulary forces us to ignore the information present in the rest of the vocabulary. We think this information can be effectively tapped using *translingual themes*. Table VII shows the improvement in MRR for AUX-COMB (which uses translingual themes) with English as the auxiliary language.³⁰ We see reasonable improvement in MRR in general, with large improvements (up to **91%**) for some language pairs. We see similar behavior with French and Hindi as the auxiliary language (Tables VIII and IX). To show the contribution of the auxiliary language model, we shade each cell in Table VII proportional to $\lambda_{\{en\}}$, the component of λ corresponding to $P_{\{en\}}$. The minimum and maximum values of $\lambda_{\{en\}}$ were 0.51 and 0.81, respectively, and the mean and median values were both 0.65.

³⁰We report the percentage improvement rather than the absolute scores for ease of comparison. The absolute scores are documented in the supplementary material. We used the mean MRR across samples and omit variances due to lack of space (e.g., the average variance was .04 for $S_{\{en\}}()$).

Table VI. Absolute Performance (in Terms of MRR) of TsuAM on $G(en)$

| MRR | bn | hi | kn | ml | mr | ta | te |
|-----|-------|-------|-------|-------|-------|-------|-------|
| bn | – | .0905 | .0811 | .0582 | .063 | .0964 | .0787 |
| hi | .0656 | – | .0591 | .0319 | .0699 | .0466 | .0478 |
| kn | .1354 | .145 | – | .0812 | .0276 | .0874 | .0811 |
| ml | .0442 | .1005 | .0653 | – | .0361 | .0504 | .0345 |
| mr | .0724 | .1967 | .0682 | .0601 | – | .0531 | .0715 |
| ta | .0507 | .0417 | .0577 | .0325 | .0289 | – | .1015 |
| te | .1009 | .089 | .0801 | .0822 | .024 | .0771 | – |

Table VII. Percentage Performance Improvement (over Baseline MRR) of AUX-COMB Using $S_{(en)}()$ (the Shading Darkness of a Cell Is Proportional to $\lambda_{(en)}$)

| %Imp | bn | hi | kn | ml | mr | ta | te |
|------|-------|-------|--------------|--------------|--------------|-------|-------|
| bn | – | 24.95 | 90.34 | 20.81 | 10.46 | 28.16 | 38.00 |
| hi | 7.89 | – | 5.71 | 24.09 | 25.02 | 25.97 | 26.14 |
| kn | 55.04 | 26.50 | – | 91.83 | 58.55 | 50.21 | 65.93 |
| ml | 12.32 | 19.08 | 37.45 | – | 17.74 | 3.93 | 36.67 |
| mr | 21.17 | 29.46 | 65.93 | 39.71 | – | 14.05 | 23.59 |
| ta | 8.46 | 9.41 | 9.67 | 4.81 | 7.81 | – | 21.35 |
| te | 29.49 | 36.94 | 81.69 | 19.53 | 33.91 | 42.98 | – |

Table VIII. Percentage Performance Improvement (over Baseline MRR) of AUX-COMB Using $S_{(fr)}()$

| %Imp | bn | hi | kn | ml | mr | ta | te |
|------|-------|-------|--------------|--------------|--------------|-------|-------|
| bn | – | 32.50 | 60.04 | 34.47 | 23.59 | 24.13 | 27.52 |
| hi | 21.37 | – | 22.92 | 31.38 | 8.63 | 18.50 | 19.71 |
| kn | 43.11 | 19.85 | – | 70.15 | 32.83 | 51.69 | 44.58 |
| ml | 22.28 | 16.10 | 55.33 | – | 11.07 | 28.29 | 39.10 |
| mr | 33.30 | 26.59 | 49.07 | 22.73 | – | 10.22 | 36.64 |
| ta | 33.63 | 11.59 | 24.97 | 21.37 | 7.51 | – | 18.22 |
| te | 20.44 | 18.15 | 59.24 | –2.52 | 24.32 | 36.07 | – |

We tried AUX-COMB with two³¹ auxiliary languages to study the impact of using more languages (Table X).³² The results are much better than when a single auxiliary language is used (we see up to **124%** improvement). For example, for mr - ml , the improvement obtained using en and fr was 39% and 22%, respectively, and using both was 83%. We see similar results for kn - te , te - mr , and so forth. We see robust performance for most of the 21 language pairs and for both directions.

Asymmetric Performance. As anticipated in Section 3.2, we see an asymmetry in performance for a single language pair; for example, MRR for te - ml is 0.3543, while MRR for ml - te is 0.2381. Since the auxiliary models also have the same property, we see that the performance improvement is also not symmetric—even if the baseline performance happens to be symmetric. For example, MRR values for ta - te are 0.25 and 0.26, while the improvements are 21% and 42%.

³¹The model allows the inclusion of any number of auxiliary languages. However, our experimental setup requires the training pairs to be present in every auxiliary language corpus, so as to accurately measure the contribution of each auxiliary language. This restriction resulted in very small training sets when using three or more auxiliary languages, for example, $|G(\{en, fr, hi\})| = 37$ for kn - ml . For this reason, we did not try with more auxiliary languages for our chosen set of language pairs.

³²Due to the very poor performance of the auxiliary language baselines on $G(en)$, they were not considered for further experiments, and we report only the results for AUX-COMB.

Table IX. Percentage Performance Improvement (over Baseline MRR) of AUX-COMB Using $S_{[hi]}$ ()

| %Imp | bn | hi | kn | ml | mr | ta | te |
|------|-------|----|--------------|--------------|--------------|-------|-------|
| bn | – | – | 61.78 | 26.08 | 23.37 | 23.79 | 25.32 |
| kn | 29.79 | – | – | 34.68 | 18.85 | 40.08 | 54.47 |
| ml | 12.22 | – | 72.33 | – | 25.58 | 44.09 | 33.28 |
| mr | 15.15 | – | 71.11 | 33.61 | – | 24.24 | 37.81 |
| ta | 19.71 | – | 24.14 | 13.63 | 19.46 | – | 34.99 |
| te | 20.71 | – | 76.78 | 19.59 | 53.45 | 54.93 | – |

Table X. Percentage Performance Improvement (over Baseline MRR) of AUX-COMB Using $S_{[en,fr]}$ ()

| %Imp | bn | hi | kn | ml | mr | ta | te |
|------|-------|-------|---------------|--------------|--------------|-------|-------|
| bn | – | 36.05 | 92.45 | 42.59 | 26.95 | 41.55 | 46.90 |
| hi | 24.96 | – | 31.77 | 28.94 | 34.75 | 25.95 | 43.81 |
| kn | 53.36 | 27.27 | – | 82.33 | 89.51 | 52.03 | 94.75 |
| ml | 13.98 | 22.83 | 51.72 | – | 23.26 | 18.77 | 68.03 |
| mr | 32.10 | 35.66 | 95.94 | 83.78 | – | 12.48 | 42.36 |
| ta | 39.64 | 17.78 | 23.22 | 15.50 | 19.12 | – | 45.39 |
| te | 33.60 | 38.21 | 124.54 | 10.24 | 70.74 | 55.37 | – |

Table XI. Examples: For Each Source kn Word, We Generate the Translation Correspondence Using TI+Cue and Using AUX-COMB (with $S_{[en,fr]}$) and Show (a) the Top-Ranked te Word, and (b) the Rank of the te Translation

| Source word | | TI+Cue | | | $S_{[en,fr,hi]}$ | | |
|-------------|--------------|---------------------|-------------|-----------------|---------------------|-------------|-----------------|
| kn word | Meaning | te word at rank 1 | Meaning | Rank of transl. | te word at rank 1 | Meaning | Rank of transl. |
| ఎలక్ట్రాన్ | electron | చక్రకర | sugar | 20 | ఎలక్ట్రాన్ | electron | 1 |
| రసవిద్య | chemistry | గ్రీక్ | Greek | 24 | శాస్త్రం | science | 3 |
| శని | saturn | సహాయము | literature | 9 | శని | saturn | 1 |
| శీలీంధ్ర | fungus | పర్యవరణ | environment | 32 | లైకన్ | lichen | 4 |
| గురుత్వ | gravitation | గురుత్వకర్షణ | gravitation | 1 | కృష్ణ | dark | 3 |
| జీవసత్వం | vitamins | వ్యాధి | disease | 55 | విటమిన్ | vitamin | 1 |
| జీవవైవిధ్య | biodiversity | పర్యవరణ | environment | 9 | పర్యవరణ | environment | 4 |
| గులామగిరి | slavery | కౌన్సిల్ | council | 2 | బనిసత్వం | slavery | 1 |

Examples from kn - te . Table XI shows examples for kn - te . For each kn word, we take the translation correspondences using TI+Cue and AUX-COMB (with $S_{[en,fr]}$ ()) and show the te word at rank 1 and the rank of the correct te translation. We found that the top-ranked terms from both approaches were topically related, but the translation was not usually at rank 1. However, AUX-COMB is able to use evidence from multiple languages and boost the probability of the translation so that it is ranked higher.

6.1.6. Further Analysis for AUX-COMB.

Small Development Sets Are Enough. We analyzed how sensitive our method was to the size of the development set used for learning the mixture weights λ . We chose a language pair (mr - te) that had a sufficiently large gold set to allow development set size ablation, and sufficiently high performance to allow both positive and negative variation. In Figure 4 (left), we see the performance of AUX-COMB for different development set sizes. We see a gradual increase in performance as development set size increases. For just 10 pairs, the performance is nearly as good as the performance for 70 pairs. The trend for te - mr was very similar. This suggests that we can learn the model with very few translation pairs, which is useful in a low-resource setting.

Both Rare and Frequent Words Do Better. We analyzed how our method performed on words with different collection frequencies. For the language pair te - mr , we plotted the

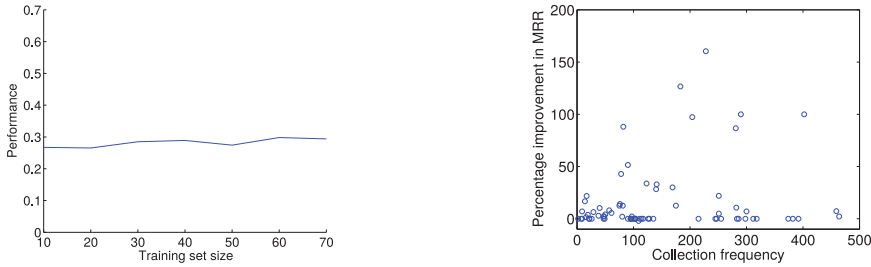


Fig. 4. Left: Performance for different development set sizes for *mr-te*. Right: Performance improvement for *te* terms with different collection frequencies, for *te-mr* with $S_{en}()$.

collection frequency of *te* words versus percent improvement in MRR (Figure 4 (right)). We observe improvement over a wide range of frequencies, suggesting that the method is suitable for both rare and frequent words. The observations were similar for *mr* terms as well. We performed similar analyses for other term properties, viz. document frequency and average document count, and observed similar behavior.

6.2. Cross-Lingual Semantic Relatedness

The standard way to evaluate measures of semantic relatedness is to compute the agreement with human-assigned scores. A set of word pairs is selected and human annotators are asked to manually assign each pair a relatedness score. Each pair is scored by at least two annotators and the scores are treated as valid only if they show good agreement. A higher agreement value indicates better consensus among annotators on the score to be assigned to different word pairs. Finally, the scores from different annotators are averaged to arrive at the final score for each word pair. This is the gold standard used to evaluate the algorithmic measures. The algorithm we want to evaluate is used to generate relatedness scores for each pair in the gold standard, and correlation between the generated scores and the gold standard scores are reported. A higher correlation is treated as an indication that the algorithm is better at mimicking human judgments of semantic relatedness. We used this experimental approach to compare the AUX-COMB-SR method with the TI+Cue-SR baseline.

Human annotation is expensive and it would be difficult to create gold standards for all 21 language pairs used in our study. Given the limitations of time and budget, we created gold standard datasets for three language pairs, such that we have one Indo-Aryan–Dravidian pair (Marathi–Malayalam), one Indo-Aryan–Indo-Aryan pair (Bengali–Marathi), and one Dravidian–Dravidian pair (Malayalam–Tamil). However, the Malayalam–Tamil dataset did not meet our quality standards and could not be used for further experiments. In the following, we report the results for Bengali–Marathi and Malayalam–Marathi.

6.2.1. New Datasets for CLSR. We took the 100 most frequent words from each language in each language pair and generated the *Cue* score [Vulić et al. 2011] for all words in the other language in the pair. For each word, we took the highest probability word in the other language and formed a pair. We combined the sets to get 200 word pairs for each language pair. This data was rated by annotators (graduate students).

Annotation Task Design. Given the diverse nature of the languages we chose to evaluate, it was difficult to find annotators well versed in both the languages in each pair. To work around this problem, we designed the annotation exercise as follows. We identified a volunteer for each language in the language pair. The volunteer was required to be a native speaker in the language assigned to him or her. The two

Table XII. Details of Two New Datasets for Semantic Relatedness: The Interannotator Agreements (in Terms of Krippendorff's α) for the Popular WordSim353 Dataset Are Given for Comparison

| Language Pair | #Word Pairs | Krippendorff's α |
|------------------|-------------|-------------------------|
| <i>bn-mr</i> | 127 | 0.846 |
| <i>ml-mr</i> | 161 | 0.882 |
| WordSim353–Set 1 | 153 | 0.667 |
| WordSim353–Set 2 | 200 | 0.471 |

volunteers for a language pair were asked to sit together and review each word pair in their assigned dataset. Each volunteer was asked to describe all known meanings of the word in his or her chosen language to the other volunteer, and vice versa. After the meanings of both words were clear to both volunteers, the volunteers were asked to stop discussing and spend some time in isolation to form an independent judgment of the semantic relatedness of the two words and record a relatedness score for that word pair. The volunteers were instructed to abstain from discussing their relatedness judgments and from revealing their scores. All annotation exercises were done under the continuous supervision of one of the authors to ensure that the instructions were strictly followed.

The relatedness score was required to be a real number between 0 and 10, where 0 indicates that the words are totally unrelated, and 10 indicates that the words are practically synonymous. We asked the annotators to consider antonyms as related (since they are features of the same concept), rather than unrelated. If a word was repeated,³³ or if an annotator could not identify the meanings of a word, the corresponding word pair was discarded from the dataset. We average the score for each word pair; this is the gold standard.

Interannotator Agreement. For each dataset, we evaluated the interannotator agreement to verify the quality of the annotation. The Cohen κ coefficient, which is frequently used for quantifying interannotator agreement, is defined for categorical annotations, and not suitable for our task where the annotations are real numbers. We instead use the Krippendorff α coefficient [Krippendorff 2004], which is suitable for content annotation tasks with real numbers [Artstein and Poesio 2008] and is widely used in content analysis tasks [Krippendorff 2012].

To use Krippendorff's α , we need to define a distance metric d between the scores r and r' given by the two annotators for a word pair, where $d(r, r')$ indicates how different r and r' are. For example, a simple distance metric could be $d(r, r') = 0$ if $r = r'$, and 1 if $r \neq r'$. This distance metric would be suitable for annotations that are class labels and where disagreement on any two classes is equally bad. For our task, which involves real numbers as scores, we would like to capture the extent of disagreement between the two scores (r, r'); for example, (2,7) is a stronger disagreement than (4,5). Also, we would like to penalize stronger disagreements more than weak disagreements like (4,5). Hence, we defined the distance metric as $d(r, r') = (r - r')^2$, which has the required properties.³⁴

Some details of the gold standard semantic relatedness datasets for Bengali–Marathi (*bn-mr*) and Malayalam–Marathi (*ml-mr*) are given in Table XII. For the

³³Due to the absence of morphological analyzers, we could not perform lemmatization on the corpora. Due to this, different forms of the same word exist as different terms in the vocabulary. For this task, we treated the variations of a word as identical to each other.

³⁴We used the Python library by Passonneau et al. [2008] for computing Krippendorff's α (available at <http://cswww.essex.ac.uk/Research/nle/arrau/Lippincott/agreement.tgz>) and modified it to incorporate our distance metric.

Table XIII. Correlation of the Semantic Relatedness Scores Generated by TI+Cue-SR and AUX-COMB-SR with the Average Relatedness Score Given by Human Annotators (the Correlations Marked with * Were Statistically Significant with $p < 0.01$)

| Correlation Coefficient | <i>bn-mr</i> | | <i>ml-mr</i> | |
|-------------------------|--------------|--------------|--------------|--------------|
| | TI+Cue-SR | AUX-COMB-SR | TI+Cue-SR | AUX-COMB-SR |
| Pearson's ρ | 0.33* | 0.43* | 0.19 | 0.35 |
| Spearman's ρ | 0.27* | 0.37* | 0.09 | 0.30* |
| Kendall's τ | 0.19* | 0.27* | 0.07 | 0.21* |

sake of comparison, we computed the Krippendorff α values on the WordSim353 dataset [Finkelstein et al. 2001] (which includes the Miller-Charles dataset [Miller and Charles 1991]), which has been widely used to evaluate algorithmic measures of semantic relatedness. Both the new datasets show strong interannotator agreement and have better interannotator agreement than the WordSim353 datasets.

6.2.2. Semantic Relatedness Evaluation. The state-of-the-art method for cross-lingual semantic relatedness is **Response-BC** introduced by Vulić and Moens [2013]. As discussed in their results, the performance of the TI+Cue method was very close to this method on all the corpora used in their experiments. For example, the average difference between the two methods for the Top 10 Accuracy (Acc_{10}) measure was less than 5.6%. Since our objective is to demonstrate the effect of auxiliary languages, rather than propose a new algorithm for cross-lingual semantic relatedness, we decided to use the TI+Cue-SR method as our baseline, since the code was already available. The TI+Cue method was also used to construct the scoring functions S_{raw}^{XY} in AUX-COMB-SR.

To evaluate the impact of using auxiliary languages, we computed the correlation between the human-assigned scores and the scores from TI+Cue-SR and AUX-COMB-SR. The results are shown in Table XIII. We see significant improvement in both linear (Pearson's ρ) and rank correlation (Spearman's ρ and Kendall's τ) when compared to the state-of-the-art baseline. The language pair *ml-mr* has languages from different families and is harder to solve, as can be seen from the extremely low correlation achieved by the baseline. AUX-COMB-SR is particularly useful in this case and causes more than a **200%** improvement in rank correlation, and more than an 84% improvement in linear correlation.

6.3. Wikipedia Title Suggestion—User Study

We performed a user study on the *WikiTSu* system for the language pair Kannada-Hindi to assess the quality of the cross-lingual titles suggested. The quality of suggestions for source words that are Wikipedia titles has been exhaustively studied in Section 6.1.5. In the user study, we focused on source words that are *not* Wikipedia titles. The Kannada Wikipedia (14K articles) is much smaller than the Hindi Wikipedia (100K articles), so we chose Kannada as the source language.

Study Methodology. We randomly selected 3,200 words from the kn corpus that were not titles and removed common verbs, adjectives, parts of names, very common nouns, and noise words—these are unlikely to be article titles in Hindi (or any other language), giving a final list of 512 words. For each kn word k , we scored the hi vocabulary and presented the top-scoring hi title h to a user, with the following instructions: Suppose a user sees k in an article and wants to know more about the concept K represented by the word k . Let H be the article corresponding to h . Score h as 1 if H is about the concept K , 0.5 if H contains information about concept K , and 0 otherwise. This exercise was performed independently by two users.

Results. For each scoring method (TI+Cue-WTS and AUX-COMB-WTS), for each k , we averaged the relevance score given by the two users and then averaged that over

Table XIV. User Study on *WikiTSu*: Average Score of Suggested Titles and User Agreement Metrics (Left) and the Weight Matrix for Weighted κ (Right)

| | TI+Cue-WTS | AUX-COMB-WTS | User 2 | | | |
|-------------------|------------|--------------|--------|---|-----|---|
| Avg. score | 0.298 | 0.360 | W | 1 | 0.5 | 0 |
| Agreement | 83% | 81% | 1 | 0 | 1 | 3 |
| Cohen's κ | 0.69 | 0.68 | 0.5 | 1 | 0 | 1 |
| Weighted κ | 0.83 | 0.81 | 0 | 3 | 1 | 0 |

all k . The results are shown in Table XIV. We see that using AUX-COMB-WTS leads to a **20% improvement** in the quality of titles. The Cohen κ agreement is good but does not take the ordering of the scores into account—a disagreement of 0 versus 1 is worse than 0 versus 0.5. We computed the weighted κ [Cohen 1968] using the weight matrix W shown in Table XIV and found very good agreement.³⁵

7. FUTURE WORK

In this article, we explored using auxiliary language corpora for two fundamental tasks in the cross-lingual domain (translation induction and semantic relatedness measurement) and one applied task (Wikipedia title (headword) suggestion). In the comparable corpora-only setting, and in the absence of any other resources, these problems are hard to solve and current approaches do not give satisfactory results. For the CC-TCI task, we demonstrated remarkable improvements in performance for 21 language pairs when using auxiliary languages. We created two new human-annotated datasets for the CC-CLSR task and demonstrated significant gains from the use of auxiliary languages. The datasets have been made publicly available to facilitate further research in this new area. For the real-world application of Wikipedia title suggestion, we built the *WikiTSu* system and have made the code and data for the system publicly available. We conducted a user study and found that using auxiliary languages helps significantly improve the quality of the output of *WikiTSu*. This study raises interesting questions regarding the effect of the number of languages, language family, and corpus characteristics and quality. The inclusion of multiword units raises new challenges. The model combination framework allows easy introduction of other cues besides auxiliary language corpora (e.g., transliteration models for names). We plan to explore these questions and ideas in future work.

ACKNOWLEDGMENTS

We thank Chaitra Shankar, Indu John, Narendra, and Balamurugan for help with the annotation, and Srivaths Ranganathan for help with the initial experiments.

REFERENCES

- Daniel Andrade, Masaaki Tsuchida, Takashi Onishi, and Kai Ishikawa. 2013. Translation acquisition using synonym sets. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 4 (2008), 555–596.
- Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: Are more languages better? In *Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 28–36.
- Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. (*ACL05*).
- Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael S. Horn, and Darren Gergle. 2012. Omnipedia: Bridging the wikipedia language gap. In *ACM Conference on Human Factors in Computing Systems*.

³⁵ W_{ab} is the penalty when a title is given the score a by user 1 and b by user 2.

- Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. 2007. Measuring semantic similarity between words using web search engines. *International World Wide Web Conference (WWW)* 7 (2007), 757–766.
- Lars Borin. 2000. You'll take the high road and I'll take the low road: Using a third language to improve bilingual word alignment. In *International Conference on Computational Linguistics*.
- Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19, 2 (1993), 263–311.
- Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70, 4 (1968), 213.
- Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Annual Meeting of the Association for Computational Linguistics*.
- Raj Dabre, Fabien Cromieres, Sadao Kurohashi, and Pushpak Bhattacharyya. 2014. Leveraging small multilingual corpora for smt using many pivot languages. In *North American Chapter of the Association for Computational Linguistics*.
- Dmitry Davidov and Ari Rappoport. 2009. Enhancement of lexical concepts using cross-lingual web mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP): Volume 2*. Association for Computational Linguistics, 852–861.
- Hervé Déjean, Éric Gaussier, and Fatiha Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *International Conference on Computational Linguistics*.
- Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2009. Improving the extraction of bilingual terminology from Wikipedia. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 5, 4 (2009), 31.
- Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *International World Wide Web Conference (WWW)*. ACM, 406–414.
- Darja Fišer and Nikola Ljubešić. 2011. Bilingual lexicon extraction from comparable corpora for closely related languages. In *Recent Advances in Natural Language Processing*.
- Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Workshop on Very Large Corpora* (1995).
- Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *International Joint Conference on Artificial Intelligence*. 6.
- Eric Gaussier, J.-M. Renders, Irina Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Annual Meeting of the Association for Computational Linguistics*. DOI: <http://dx.doi.org/10.3115/1218955.1219022>
- Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.
- Samer Hassan, Carmen Banea, and Rada Mihalcea. 2012. Measuring semantic relatedness using multilingual representations. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval'12)*. Association for Computational Linguistics, Stroudsburg, PA, 20–29.
- Samer Hassan and Rada Mihalcea. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Conference on Empirical Methods in Natural Language Processing*. 10.
- Ann Irvine and Chris Callison-Burch. 2013. Supervised bilingual lexicon induction with multiple monolingual signals. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Azniah Ismail and Suresh Manandhar. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In *International Conference on Computational Linguistics*.
- Glen Jeh and Jennifer Widom. 2002. SimRank: A measure of structural-context similarity. In *Conference on Knowledge Discovery and Data Mining*. 6.
- Heng Ji. 2009. Mining name translations from comparable corpora by creating bilingual information networks. In *Workshop on Building and Using Comparable Corpora*. 4.
- Hiroyuki Kaji, Shin'ichi Tamamura, and Dashtseren Erdenebat. 2008. Automatic construction of a Japanese-Chinese dictionary via English. In *Language Resources and Evaluation Conference*.
- Mitesh M. Khapra, A. Kumaran, and Pushpak Bhattacharyya. 2010. Everybody loves a rich cousin: An empirical study of transliteration through bridge languages. In *Human Language Technology Conference*. 9.

- Woosung Kim and Sanjeev Khudanpur. 2004. Lexical triggers and latent semantic analysis for cross-lingual language model adaptation. *ACM Transactions on Asian Language Information Processing (TALIP)* 3, 2 (June 2004), 94–112.
- Alexandre Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2013. Toward statistical machine translation without parallel corpora. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *ACL Workshop on Unsupervised Lexical Acquisition*.
- Klaus Krippendorff. 2004. Content analysis: An introduction to its methodology. Sage.
- Klaus Krippendorff. 2012. *Content Analysis: An Introduction to Its Methodology*. Sage.
- Shankar Kumar, Franz Josef Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *Conference on Empirical Methods in Natural Language Processing: Conference on Natural Language Learning*.
- Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *International Conference on Computational Linguistics*.
- Florian Laws, Lukas Michelbacher, Beate Dorow, Christian Scheible, Ulrich Heid, and Hinrich Schütze. 2010. A linguistically grounded graph model for bilingual lexicon extraction. In *International Conference on Computational Linguistics*.
- Bo Li and Eric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *International Conference on Computational Linguistics*.
- Lishuang Li, Peng Wang, Degen Huang, and Lian Zhao. 2011. Mining English-Chinese named entity pairs from comparable corpora. *ACM Transactions on Asian Language Information Processing* 10, 4 (2011), 19.
- Gideon S. Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *North American Chapter of the Association for Computational Linguistics*. 8. DOI: <http://dx.doi.org/10.3115/1073336.1073356>
- Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Michael Skinner, and Jeff Bilmes. 2009. Compiling a massive, multilingual dictionary via probabilistic inference. In *Annual Meeting of the Association for Computational Linguistics*.
- George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6, 1 (1991), 1–28.
- David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Conference on Empirical Methods in Natural Language Processing*.
- Emmanuel Morin, Batrice Daille, Koichi Takeuchi, and Kyo Kageura. 2008. Brains, not brawn: The use of \smart\ comparable corpora in bilingual terminology mining. *ACM Transactions on Speech and Language Processing* 7, 1 (2008), 1.
- Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelRelate! A joint multilingual approach to computing semantic relatedness. In *AAAI*. 22–26.
- Rebecca J. Passonneau, Tae Yano, Tom Lippincott, and Judith Klavans. 2008. Relation between agreement measures on human labeling and machine learning performance: Results from an art history image indexing domain. *Computational Linguistics for Metadata Building* (2008), 49.
- Richard R. Picard and R. Dennis Cook. 1984. Cross-validation of regression models. *Journal of the American Statistical Association* 79, 387 (1984), 575–583.
- Emmanuel Prochasson and Pascale Fung. 2011. Rare word translation extraction from aligned comparable documents. In *Human Language Technology Conference*.
- Longhua Qian, Hongling Wang, Guodong Zhou, and Qiaoming Zhu. 2012. Bilingual lexicon construction from comparable corpora via dependency mapping. In *International Conference on Computational Linguistics*.
- Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Annual Meeting of the Association for Computational Linguistics*. 3.
- Reinhard Rapp. 1996. *Die Berechnung von Assoziationen: Ein Korpuslinguistischer Ansatz*. Vol. 16. Georg Olms Verlag.
- Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Annual Meeting of the Association for Computational Linguistics*. 8. DOI: <http://dx.doi.org/10.3115/1034678.1034756>
- Reinhard Rapp, Serge Sharoff, and Bogdan Babych. 2012. Identifying word translations from comparable documents without a seed lexicon. In *Language Resources and Evaluation Conference*. 460–466.
- Sujith Ravi. 2013. Scalable decipherment for machine translation via hash sampling. In *Annual Meeting of the Association for Computational Linguistics (1)*. 362–371.

- Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 12–21.
- Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM* 8, 10 (1965), 627–633.
- Raphaël Rubino and Georges Linarès. 2011. A multi-view approach for term translation spotting. In *Computational Linguistics and Intelligent Text Processing*.
- Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *International Conference on Computational Linguistics*. 7. DOI : <http://dx.doi.org/10.3115/1118853.1118879>
- Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. 2013. Overviewing important aspects of the last twenty years of research in comparable corpora. In *Workshop on Building and Using Comparable Corpora (BUCC)*. Springer, 1–17.
- Daphna Shezaf and Ari Rappoport. 2010. Bilingual lexicon generation using non-aligned signatures. In *Annual Meeting of the Association for Computational Linguistics*.
- Fangzhong Su and Bogdan Babych. 2012. Development and application of a cross-language document comparability metric. In *Language Resources and Evaluation Conference*.
- Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *Conference on Empirical Methods in Natural Language Processing: Conference on Natural Language Learning*.
- Takashi Tsunakawa, Naoaki Okazaki, and Jun ichi Tsujii. 2008. Building bilingual lexicons using lexical translation probabilities via pivot languages. In *Language Resources and Evaluation Conference*.
- Peter Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. *European Conference on Machine Learning*. Springer, 491–502.
- Raghavendra Udupa and Mitesh Khapra. 2010. Improving the multilingual user experience of Wikipedia using cross-language name search. In *Human Language Technology Conference*. 9.
- Raghavendra Udupa, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. 2009. MINT: A method for effective and scalable mining of named entity transliterations from large comparable corpora. In *Conference of the European Chapter of the Association for Computational Linguistics*. 9.
- Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies: North American Chapter of the Association for Computational Linguistics*.
- Ellen M. Voorhees. 1999. The TREC-8 question answering track report. In *Text REtrieval Conference*. Vol. 99, 77–82.
- Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. <http://www.aclweb.org/anthology/P11-2084>.
- Ivan Vulić and Marie-Francine Moens. 2013. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation* 21, 3 (2007), 165–181. DOI : <http://dx.doi.org/10.1007/s10590-008-9041-6>

Received October 2015; revised March 2016; accepted December 2016