Journal of Biomedical Informatics 85 (2018) 56–67

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Methodological Review

Development of machine translation technology for assisting health communication: A systematic review

Kristin N. Dew[a,1], Anne M. Turner[b,1,*], Yong K. Choi[c], Alyssa Bosold[d], Katrin Kirchhoff[e]

[a] University of Washington, Dept. of Human Centered Design & Engineering, Northwest Center for Public Health Practice, 1107 NE 45th Street, Suite 400, Seattle, WA 98105, USA
[b] University of Washington, Dept. of Health Services and Dept. Biomedical Informatics and Medical Education, Northwest Center for Public Health Practice, 1107 NE 45th Street, Suite 400, Seattle, WA 98105, USA
[c] University of Washington, Dept. of Biomedical Informatics and Medical Education, Seattle, WA, USA
[d] University of Washington, Dept. of Health Services, Seattle, WA, USA
[e] University of Washington, Dept. of Electrical Engineering, Seattle, WA, USA

ARTICLE INFO

Keywords:
Public health
Public health informatics
Natural language processing
Health communication
Health literacy

ABSTRACT

Objectives: To (1) characterize how machine translation (MT) is being developed to overcome language barriers in health settings; and (2) based on evaluations presented in the literature, determine which MT approaches show evidence of promise and what steps need to be taken to encourage adoption of MT technologies in health settings.

Materials & methods: We performed a systematic literature search covering 2006–2016 in major health, engineering, and computer science databases. After removing duplicates, two levels of screening identified 27 articles for full text review and analysis. Our review and qualitative analysis covered application setting, target users, underlying technology, whether MT was used in isolation or in combination with human editing, languages tested, evaluation methods, findings, and identified gaps.

Results: Of 27 studies, a majority focused on MT systems for use in clinical settings (n = 18), and eight of these involved speech-based MT systems for facilitating patient-provider communications. Text-based MT systems (n = 19) aimed at generating a range of multilingual health materials. Almost a third of all studies (n = 8) pointed to MT's potential as a starting point before human input. Studies employed a variety of human and automatic MT evaluation methods. In comparison studies, statistical machine translation (SMT) systems were more accurate than rule-based systems when large corpora were available. For a variety of systems, performance was best for translations of simple, less technical sentences and from English to Western European languages. Only one system has been fully deployed.

Conclusions: MT is currently being developed primarily through pilot studies to improve multilingual communication in health settings and to increase access to health resources for a variety of languages. However, continued concerns about accuracy limit the deployment of MT systems in these settings. The variety of piloted systems and the lack of shared evaluation criteria will likely continue to impede adoption in health settings, where excellent accuracy and a strong evidence base are critical. Greater translation accuracy and use of standard evaluation criteria would encourage deployment of MT into health settings. For now, the literature points to using MT in health communication as an initial step to be followed by human correction.

1. Objective

In the US, the increasing population of non-English speakers and related regulatory requirements are driving a need for translation in health communications. Machine translation (MT) is improving in quality, is widely available, and is being taken up in business applications and translation services. However, whether current MT approaches are adequate for the specific needs of health care settings is unclear. In this paper, we review the literature over the period from 2006–2016 to describe the extent to which machine translation has been investigated in the context of health settings and whether current MT technologies hold promise for improving language communication

* Corresponding author.
E-mail addresses: kndew@uw.edu (K.N. Dew), amturner@uw.edu (A.M. Turner).
[1] Kristin N. Dew and Anne M. Turner contributed equally to this manuscript and worked closely in its preparation.

https://doi.org/10.1016/j.jbi.2018.07.018
Received 19 August 2017; Received in revised form 16 July 2018; Accepted 18 July 2018
Available online 19 July 2018
1532-0464/ © 2018 Published by Elsevier Inc.

in healthcare. This qualitative review was guided by two main research questions: (1) In what ways is MT currently being developed to overcome linguistic barriers in health settings? and (2) based on evaluations presented in the literature, which MT approaches show evidence of promise for adoption in health settings?

2. Background & significance

Providing language-appropriate access to health care services is of increasing concern, both in the US and internationally. In the US, where over 300 languages are currently spoken, more than 21% of the population over five years of age speaks English as a second language [1,2]. Among those who use English as a second language, more than 25 million have limited English proficiency (LEP), or the limited ability to speak, read, write, or understand English [1]. Due to language barriers, individuals with LEP face difficulties accessing health education resources and healthcare. LEP status has been linked to both higher health care costs and greater health disparities – such as less preventive health screening – compared to individuals who are not native English speakers but who are proficient in English [1,3–9].

To help address this health gap, US healthcare service providers and insurers have come under regulatory pressure in recent years to provide language appropriate materials and services for individuals who do not understand English [8]. Language-appropriate access to federally funded services has been mandated under Title VI of the Civil Rights Act of 1964 and its associated regulations, such as the National Standards for Culturally and Linguistically Appropriate Services in Health and Health Care (CLAS standards), were established to improve healthcare by providing a framework for organizations that serve language diverse populations [9,10–12]. The Patient Protection and Affordable Care Act of 2010 (ACA) further expanded civil rights provisions for language access to other parts of the health care sector, such as insurers, exchanges, and other entities whose only source of federal funding is through insurance contracts [8,13]. Despite these regulations, healthcare providers face challenges in generating translated resources and language-appropriate care due to the time, work burden, and expense of translating and interpreting services, whether using limited in-house staff or contracting with professional vendors. These constraints are particularly pertinent for smaller organizations with limited resources and few multilingual staff [14].

The need for language-appropriate healthcare services has also grown internationally. With increased globalization, there is a greater need for tools that serve the diverse language needs of tourists, immigrants, refugees, and expatriates. The past two decades have brought reduced barriers to the movement of people, equipment, and services, greater availability of online information, and a concerted effort by governments and private sector actors alike to attract revenue via their healthcare systems [15].

Freely available MT offers a promising, low-cost, and efficient solution for language translation. For users who typically rely on translation vendors, combining MT with human correction can yield a comparable quality of translation in less time for as little as 5% of the cost of a human translation [14]. MT is now commonly used by translation vendors and in business applications, where MT is either used as the sole system or in combination with other systems or human post-editing [16]. Approaches like statistical machine translation (SMT), which combine large-scale data resources and state-of-the-art machine learning, have made dramatic technical progress over the past decade, with dozens of translation tools for both speech and text now freely available online and on mobile devices (e.g. Google Translate, Bing Translator).

The current predominant approach to machine translation is SMT, which is increasingly supplanting older approaches like rule-based translation (RBT) and example-based translation (EBMT). Under the statistical approach, translation models are trained automatically from large parallel corpora, i.e. texts in the source language paired with translations in the target language [17,18]. Statistical models tend to improve over time, as more parallel and monolingual texts become available; thus, for high-resource languages and matched training and test data, often accurate and fluent machine translations can be produced. On the other hand, SMT does not explicitly model deeper semantics or contextual knowledge, and often generates unacceptable results for under-resourced languages, styles, or domains. RBT, based on principles of natural language processing (NLP), requires the development of algorithms that recognize and process the syntax of source language; and uses linguistic rules to transfer the meaning to the target language [19]. Outside of the health domain, RBT is most useful with less common language pairs, when parallel bilingual corpora are limited. EBMT utilizes stored databases of manually produced translations, usually to aid human translators. Freely available translation engines like Google Translate or Microsoft Translator are based on SMT; however, hybrid approaches combining elements of both SMT and RBT/EMBT still play a role in commercial MT systems. In practice, most commercial MT vendors use a combination of approaches.

Whereas previous SMT systems used a variety of specialized models to address lexical translation, word reordering, etc., more recent models rely on complex ("deep") neural networks to perform the mapping from source to target language, which have resulted in a major leap in performance [20]. A number of recent studies have reported major improvements of neural machine translation (NMT) over earlier SMT systems [21,22]; however, it has also been noted that neural MT performs worse when training data is extremely limited, e.g., in the case of low-resource languages or specialized domains such as medicine. Most major MT providers are now shifting toward neural MT.

In addition to text translation, improvements have been made to speech translation, i.e. the automatic translation of speech rather than text input. In speech translation, a speech recognizer decodes the speech signal into text, and a MT system processes the text into another language that is then converted into speech. As is the case with MT for text, the availability of more advanced statistical models, larger data resources, and more powerful computer hardware has contributed to rapid growth in this field.

Commercial MT has become a multi-billion-dollar market. MT is used regularly in controlled domains (e.g., to translate technical manuals) and as a first pass solution, producing output that is then fine-tuned or post-corrected by human translators. A large number of enterprises as well as smaller software developers increasingly rely on MT for localization and analysis of documents in other languages. Compared to its use in the business setting, however, the potential of MT in clinical and other health related settings is less understood. This is in part due to the need for the utmost accuracy in health and safety-critical settings. Even in well-sourced language pairs like Spanish-English, it is unclear whether MT is adequate for providing comprehensive professional translations, especially when the content pertains to specialized domains (e.g., medicine), which have their own vocabularies and forms of communication, such as health records, reports, and discharge notes.

3. Materials & methods

We used a systematic review process to (1) investigate the degree to which MT technology is being developed for use for health settings; and (2) to synthesize evidence concerning where MT might be most appropriately used to facilitate communication and provide access to translated materials.

3.1. Inclusion/exclusion criteria and protocol

We limited our review to peer-reviewed articles (including other systematic reviews) that were published in English. Because MT technology has evolved quickly, we limited the search to the past 10 years, with publication dates 2006 through 2016. Inclusion criteria required

that articles:

- report on the use of MT for facilitating health communication across different human languages automatically;
- be empirical, excluding work such as editorials, commentaries, and white papers describing a technology but lacking evaluation or other empirical support;
- rely on MT technologies, not translation memories or dictionaries.

To maintain a focus on health communications across languages, we excluded articles that were primarily focused on the technical implementations of MT methods, models, or techniques (such as domain adaptation methods), and articles about MT used in service of data extraction, modeling, and other non-communications applications. We excluded works in progress and reports of preliminary results that were published as a full submission at a later date.

Our review process was based on our adaptation of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2009 checklist [23]. PRISMA is oriented toward clinical trials, which follow a common research paradigm and have standardized study designs to support universal interpretations of the results and to build an evidence base. We modified the PRISMA protocol to accommodate the range of study methodologies we found, which included qualitative and mixed-methods studies. As such, we reviewed the findings in light of the methods and evaluation metrics used, instead of the standardized outcome variables one would find in a traditional systematic review of controlled trials. Similarly, some PRISMA items, such as the risk of bias assessment, do not map to mixed methods or qualitative studies and were not included. As detailed below, our protocol involved four stages: a pilot search, a systematic search, study selection, and a qualitative review of the findings.

3.2. Pilot search

Given the interdisciplinary nature of this topic and the absence of standardized terms, in July 2016, we performed pilot searches across dozens of potential databases in health, computer science, and engineering. The pilot indicated which search terms would return the most articles of interest. Databases and search terms had to return relevant results to be included. We excluded the Cochrane Library and Association for Computational Linguistics (ACL) Anthology databases from the systematic search after they returned no results. We searched the arXiv.org database where late breaking research may be published, but because the articles have not yet been peer reviewed, we excluded this database from the systematic search described below.

3.3. Article identification

In February 2017, we performed a systematic search of English-language articles in selected databases in the health and engineering domains. The databases covered were PubMed, PubMed Central, Cumulative Index to Nursing and Allied Health Literature (CINAHL), EMBASE, Association for Computing Machinery (ACM) Digital Library, Institute of Electrical and Electronics Engineers (IEEE) Xplore, Machine Translation (MT) Archive, Compendex, and Inspec (See Table 1). Where a complete full text search returned a large number of irrelevant results (200+), the search was narrowed to ensure the relevance of the returned records. This was the case for IEEE Xplore, where we performed a metadata only search, looking only for terms in the article's abstract and bibliographic citation data. The MT Archive is an electronic repository of MT related publications that does not provide a search function. Therefore, we included articles under relevant subject groups ("medicine and health" and "medical texts") listed in the "applications" track.

Based on our pilot results (see Section 3.2 above), we used the following combinations of search terms in our queries: ("Machine translation" OR "automated translation" OR "automatic translation") AND ("health" OR "medicine" OR "nursing" OR "clinical"). See Appendix 1 for detailed keyword descriptions and search results.

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2018.07.018.

Of note, our search terms returned many articles that involved the use of computer technology in translational sciences and not *human* language translation. In our initial search, we tried to combine the MESH term "Translating" with our search keywords to isolate the language translation articles, but the number of resulting articles were too few. We noticed that many of the relevant articles were not labeled under that MESH term. We also attempted to negate terms such as "biology" and "genetics," but again, the results were too restrictive, so we decided not to apply negation and exclude translational sciences articles manually.

Table 1
Included databases.

Database	Description	Citations returned[a]	# Included in the final review[a]
PubMed & PubMed Central	PubMed & PubMed Central cover the fields of biomedicine, health care, nursing, dentistry, veterinary medicine, biomedical aspects of technology, life sciences, and social sciences	291	11
Cumulative Index to Nursing and Allied Health Literature (CINAHL)	CINAHL includes literature on nursing, biomedicine, alternative and complementary medicine, consumer health, and other allied health topics	10	3
Embase	Embase covers the biomedical field including but not limited to pharmacology, pharmaceutical science, clinical research, allied health topics, and veterinary health	13	0
Association for Computing Machinery (ACM) Digital Library	This database is focused on the field of computing and information technology. It includes all journals, conference proceedings, magazines, newsletters and books published by ACM as well as by other selected publishers	15	0
Institute of Electrical and Electronics Engineers (IEEE) Xplore	IEEE Xplore covers the fields of electrical engineering, computer science, and electronics	13	3
Machine Translation (MT) Archive	MT Archive covers topics in machine translation, computer translation, computer translation systems, and computer-based translation tools. It includes articles, books, papers, and conference proceedings	69	8
Compendex & Inspec	Compendex and Inspec are an engineering literature database covering the fields of applied physics, computing, control, bioengineering and biotechnology, food science and technology, materials science, instrumentation including medical devices, and nanotechnology	110	6

Keyword Search Terms Used: ("Machine translation" OR "automated translation" OR "automatic translation") AND ("health" OR "medicine" OR "nursing" OR "clinical")

[a] Numbers include duplicate articles across databases.

3.4. Screening & eligibility

The systematic search described above returned 507 records, excluding duplicates from within each database. We excluded other duplicates across databases (n = 75) and citations of entire workshop and conference proceedings absent a specific article citation (n = 17). Once we had identified the records of interest (n = 415), two authors performed manual review of each citation and abstract, selecting articles for full-text review using the inclusion/exclusion criteria described above. The first author reviewed all abstracts, and the second, third, and fourth authors each reviewed a portion of the abstracts to decide on inclusion or exclusion from the full text review. Where there was disagreement, articles were reviewed in full text.

This left us with 59 articles for full text, or second-round eligibility review. In addition, we scanned reference lists to identify additional primary studies. Three articles were found in this manner [24–26], leaving us with 62 articles for the second-round eligibility review.

For the second-round review, two authors conducted a blinded review of each of the 62 articles and marked it for inclusion or exclusion. In cases where the two authors disagreed or were unsure, the article was discussed by four authors until consensus on its inclusion or exclusion was reached. We excluded works in progress and reports of preliminary results that were published as a full submission later (e.g. results from Starlander et al. on two-way medical speech translation [27], Turner et al. on English-Chinese public health translation [28], and Seligman et al. on a patient-provider speech translation tool [29]. We also excluded works found to lack empirical evidence for their claims during the full text review, i.e. works that claimed a technology's potential usefulness but did not support this claim with an evaluation. Twenty-six articles remained after the second review [14,21,24–26,30–50], with one additional article identified by further manual review at this stage [51], bringing the total to 27. See Fig. 1 for the PRISMA-based flow chart for study selection.

3.5. Full text review & qualitative analysis

To answer our review questions, we assessed the full text of each of the 27 included articles [14,21,24–26,30–51], across common dimensions. To understand how MT is being developed to improve health communications, we characterized the types of applications inductively based on dimensions present in all papers and their pertinence to health settings: their stated purpose and setting; their target languages; and

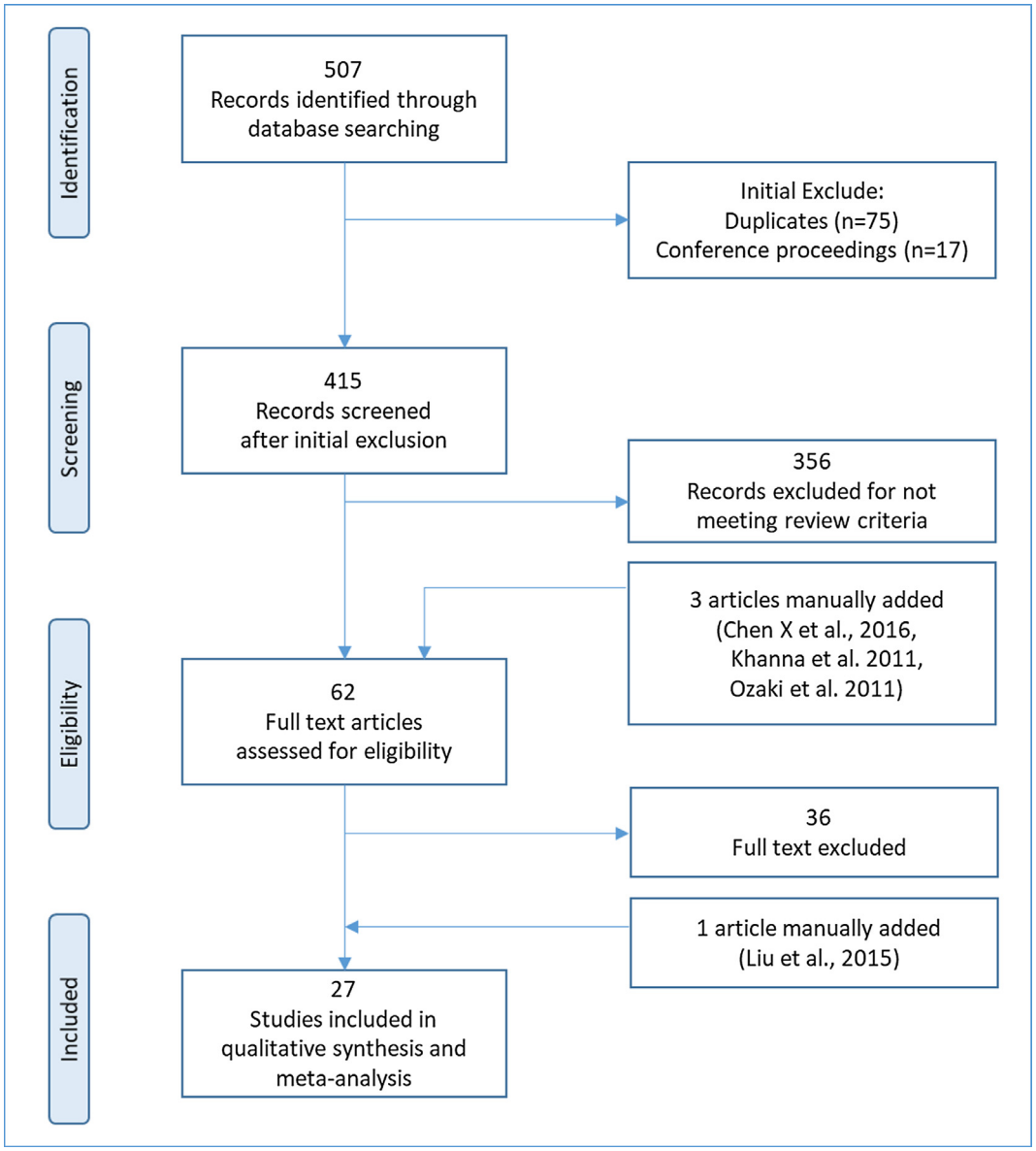

Fig. 1. Flow chart of review process.

their underlying MT approach. To determine which approaches showed evidence of promise and which further steps need to be taken, we identified the system used in the study, the type of evaluation performed, and the authors' suggestions for future work. We summarized and drew our conclusions based on thematic qualitative analysis [52].

4. Results

4.1. Search results

The themes that emerged are presented in the sections that follow. To answer our first main research question about how machine translation (MT) is being developed, we first characterized the translation technologies by their stated purpose and setting, source and target languages, and MT approach (whether SMT, RBT, or hybrid). This is critical to understanding which health communications problems researchers are targeting. We then examined the approaches to MT development along sub-dimensions that emerged from the articles themselves: which kind of underlying engine was used (freely available or custom) and whether it was designed to translate speech or text. This closer look at how the underlying technology works helps characterize the MT methods currently being used and whether there are differences in outcome when compared. To further answer our second main research question regarding which MT approaches show evidence of promise for adoption in health settings, we analyzed the evaluation methods and findings, as well as current gaps authors pointed out for future work. For a complete list of articles and their results, see Appendix 2.

4.2. What were the technologies' stated purpose and setting?

We looked first at how authors described their technologies' purpose and setting to see what health communications challenges MT is being developed to address (See Table 2). The target users for MT applications were primarily patients who do not speak the dominant language of their medical environment, such as LEP individuals in the US [38] and English speakers in Japan and China [24].

Table 2
Overview of MT purpose and setting.

Purpose	Articles
Health Education	
Public Health	Blench [50] Khanna [26] Kirchhoff [45] Mandel [44] Turner [14] Turner [31] Chen [25] Costa-Jussa [43]
Consumer Health	Pozo [46]
Biomedical Text	Zeng-Treitler [49] Wu [21] Anazawa [41] Anazawa [42] Dwivedi [48] Liu [51] Taylor [47]
Clinical Communication	Bouillon [39] Bouillon [40] Ehsani, [37] Starlander [30] Fukushima [33] Ozaki [24] Soller [38] Patil [36] Seligman [32] Shin, S [35] Muhaxov [34]
Other (Surveillance)	Blench [50]

Of the 27 articles, only one article described a system that had been fully deployed for long-term use [50], a global disease surveillance system which uses MT for translation of health related news articles from other languages to English. All other articles described pilot systems or evaluations of MT for potential use in health settings or investigated current translation processes for potential future implementation. Eight articles described speech translation systems involving MT [24,30,32,35,37–39,40]; the remaining 19 articles involved MT of text only [14,21,25,26,31,33,34,36,41–46,48–51]. Eleven articles described that the goal of the MT system was to improve communication between health care providers/staff and patients in clinical situations in which there was a language incongruence [24,30,32,33–40].

Several studies were early pilots to test the accuracy of the MT system for potential clinical use. Although the intended use of these systems was to improve provider/staff communication with patients who did not speak the same language in clinics or hospitals, only one study involved evaluations which took place in a clinical setting [32]. The MedSLT, Converser, S-MINDS, and humanoid robot project described by Shin et al. [35], are speech systems aimed at eliciting chief complaints from patients and assisting medical staff in accurate triage and diagnosis [30,32,35,37–40]. Similarly, Petit Translator and the questionnaire system described by Fukushima et al. [33] were designed to facilitate communication between hospital staff and patients, but through text rather than speech translation [24,33]. Only one clinical application was designed for remote communication or distance care; the system linked limited Chinese proficiency users in rural Western China with doctors in urban areas and generated a pdf of patient-provider exchanges to serve as an electronic health record (EHR) [34].

Many of the studies were also aimed at developing MT tools to increase access to multilingual health education materials including public health materials [14,25,26,31,44,45], biomedical literature [21,41,42,47–49,51], and consumer health topics [46]. For LEP populations in the US, Turner et al. and Kirchhoff et al. performed a series of studies investigating the adaptation of generic SMT to produce accurate and inexpensive multilingual public health promotion materials [14,31,44,45]. Three additional studies examined translation quality in the context of multilingual health and performed comparative evaluations of Google Translate and other methods of translation for English-Spanish, English-Chinese, and Catalan-Spanish materials [25,26,43]. Finally, one study tested a mobile, multilingual diet management tool for translating foods and nutritional information when travelling, for the purpose of helping patients make nutritional choices in line with their dietary restrictions [46]. Other MT applications aim to improve access to the published medical literature [41,42], research materials [47], biomedical texts [21], homeopathic and traditional Indian medical practices [48], and EHRs in non-English languages [49,51]. Targeted users of these systems included nursing students [41,42], health care providers [48], and patients [21,47,49,51].

In the only international-scale field deployment of MT technology, the Global Public Health Information Network (GPHIN) pulls current online disease surveillance reports, and then translates, analyzes, and disseminates early warnings to relevant agencies and stakeholders in population health [50].

4.3. What were the technologies' targeted languages?

We then looked at which language pairs were being tested, both to characterize which language needs are being targeted and to see if there were relationships between language pairs and translation quality. This is because the underlying language resources matter, particularly for SMT systems, which perform better with language pairs that are well represented online. As shown in Table 3, the studies involved

Table 3
Targeted languages tested.

Article	Language pairs
Bouillon [39]	English-Spanish
Bouillon [40]	English-Arabic
Blench [50]	English-Chinese(simplified), English-Chinese(Traditional), English-Farsi, English-French, English-Russian, English-Portuguese, English-Spanish
Ehsani, [37]	English-Spanish
Starlander [30]	English-Spanish
Zeng-Treitler [49]	English-Spanish, English-Chinese, English-Russian, English-Korean
Fukushima [33]	Japanese-Chinese
Khanna [26]	English-Spanish
Kirchhoff [45]	English-Spanish
Ozaki [24]	Japanese-Chinese
Pozo [46]	English-Spanish
Wu [21]	English-French, English-Spanish, English-German, English-Hungarian, English-Turkish, English-Polish
Soller [38]	English-Spanish
Anazawa [41]	English-Japanese
Anazawa [42]	English-Japanese
Dwivedi [48]	English-Hindi
Mandel [44]	n/a[a]
Patil [36][b]	English- 8 Western European languages, English- 5 Eastern European languages, English- 11 Asian languages, English - 2 African languages
Liu [51]	English-Spanish
Shin, S [35]	English-Korean
Seligman [32]	English-Spanish
Taylor [47]	English-Arabic, English-Bulgarian, English-Chinese(simple), English-Czech, English-Danish, English-Dutch, English-French, English-German, English-Hebrew, English-Hungarian, English-Italian, English-Japanese, English-Korean, English-Norwegian(Bokmal), English-Polish, English-Portuguese, English-Romanian, English-Russian, English-Serbian, English-Spanish, English-Swedish, English-Thai, English-Turkish, English-Ukrainian
Turner [31]	English-Chinese
Chen [25]	English-Spanish, English-Chinese
Costa-Jussa [43]	Spanish-Catalan
Muhaxov [34]	Kazak-Chinese

[a] n/a: not applicable.
[b] Exact languages were not specified.

translation from a variety of source languages to a large number of different languages.

Nine of the MT systems were designed to support bidirectional communication between individuals [24,30,32,33,37,38,39,43]. Seven of the nine involved speech-to-speech systems designed for translation between physicians and patients [24,30,32,37,38,39]. By far, the most common language pair evaluated was English-Spanish (n = 11). Other language pairs included Japanese-English (n = 2), Chinese-English (n = 2), Chinese-Japanese (n = 1), Korean-English (n = 1), Arabic-English (n = 1), Uyghur/Kazakh-Chinese (n = 1), Catalan-Spanish (n = 1), and Hindi-English (n = 1). Five studies tested MT applications with four or more language pairs which included European, Asian, and African languages [21,33,36,49,50]. Two papers described workflow studies of translation processes to evaluate the potential use of MT in health settings, but did not target a specific language pair [14,44].

4.4. What was the underlying MT approach?

We lastly characterized MT technologies for health communications by their underlying approach, which highlights the different ways in which MT methods can be assembled into a translation system, each with implications for the system's performance. SMT models improve over time as more instances of the texts on which they train become available online. However, it tends to be less accurate for under-resourced languages and terminology domains such as medicine, so some health translation researchers rely on RBT or hybrid approaches which combine the two.

4.4.1. SMT vs. RBT vs. hybrid

Thirteen articles focused on implementations of freely available online SMT engines, such as Google Translate [21,25,26,31,36,38,41–43,45–47,49], 15 involved hybrid engines [21,24,30–35,37–40,46–48,50]; one tested five engines, including both freely available SMT and a hybrid system [51]. Of the hybrid engines described in detail, two relied on Language Grid, an intelligence platform that can combine dictionaries and MT tools to make custom language services and higher-quality translations [24,33]; four used the publicly available Moses SMT toolkit [21,34,46,51]; and eight used rule-based systems [30,35,37–40,43,48].

Of all the systems implemented, about half reported on the use of human analysis and correction to improve the final output of translations. This included post-editing repair [14,31,45], clarifying output with image [33], human analysis for relevance and correction [50], and MT engine training and improvement by supervised machine learning [34].

4.4.2. Text vs. speech

While the majority of implementations targeted text translation, several applications were designed for real-time speech interaction using dedicated devices in clinical settings [24,30,32,35,37–40].

4.4.2.1. Text translations. Of the 19 studies that investigated or described text-based applications for health materials, 13 utilized SMT (See Table 4). Google Translate was the most commonly evaluated SMT system, and English to Spanish was the most commonly evaluated language pair (see Section 4.5 below for evaluation results).

Only two text systems were based exclusively on RBT [48,49], with the remainder being hybrid systems or not clearly described.

4.4.2.2. Speech-to-speech translations. The intended use of these systems is to facilitate communication between providers and patients during medical encounters. Speech translations involve an initial speech to text converter, followed by machine translation and then conversion of text to speech. Eight of the 27 final articles reported on the use of MT for speech-to speech translation [24,26,30,32,35,37–40]. Studies included a variety of languages (English to Spanish, Chinese, Korean, and Arabic), and English to Spanish was again the most common language pair. In contrast to the text translation systems, the speech systems more frequently utilized a RBT [26,30,32,35,38–40] or hybrid [24,37] approach.

4.5. How was the technology evaluated?

We now turn to the second research question: Which MT approaches show evidence of promise for adoption in health settings? To answer this question, we first review the papers' evaluations and their outcomes, then the future work or gaps outlined by their authors.

About half of the studies (n = 14) used a mixed method approach to study both design and evaluation components [24,30,32–34,37–40,45,46,48,50,51]. This is common in engineering fields where a system design and rationale are presented and then evaluated using quantitative or qualitative methods such as user surveys and task-based tests. Eight studies used quantitative evaluation methods exclusively [21,25,26,35,36,42,43,47]. There were no randomized trials or cohort studies. Mandel et al. used qualitative methods based on in-depth interviews and workflow mapping in a formative study to inform MT system design [44], but did not evaluate a system in development.

All but two of the studies performed an evaluation involving a MT system, but the nature and strength of the evaluation findings varied dramatically across studies. Furthermore, only two were evaluated in their respective field settings [32,50]. Even then, the lack of strong

Table 4
Summary of Machine Translation (MT) system evaluations.

Article	What was evaluated	MT Approach	Parameter Measured	Evaluation Instrument	Key Findings
TEXT SYSTEMS					
Blench [50]	Performance of GPHIN[a] surveillance system compared to actual outbreaks	Best of breed	Efficiency, effectiveness, and timeliness of system	Details of evaluation not provided	Performed effectively and efficiently when compared to actual outbreaks.
Zeng-Treitler [49]	Comparison of Babel Fish vs human translation of sentences from medical records	Rule based translation (RBT)	Translation quality, understandability, and accuracy	Human rating using 3-point Likert scale	Performed best with English-Spanish but error rate unacceptably high for clinical use. Suggest narrowing domain, standardizing syntax, and use of medical dictionaries.
Fukushima [33]	Creation of interview sheet using SMT (Language Grid) vs parallel corpora	Statistical machine translation (SMT)	Usability, time to task completion, Adequacy, Accuracy	Human rating using Walker's adequacy evaluation method	Parallel corpora more accurate and efficient than SMT. Suggest improving accuracy through including images.
Khanna [26]	Comparison of Google Translate (GT) and human translation of text from patient education material (warfarin)	SMT	Fluency, adequacy, meaning and severity of error, and preference	Human rating and automated rating using METOR[b]	GT rated worse than human translation in terms of fluency and number of errors. Performed best with simple sentence structure and full sentences.
Kirchhoff [45]	Comparison of GT plus human post-editing vs human translation of health education materials	SMT	Adequacy, fluency, linguistic error analysis, and error rate	Human rating using linguistic error analysis	MT plus human post-editing of English to Spanish translations rated equivalent to manual translations.
Pozo [46]	Comparison of modified Moses system with GT for diet profile	SMT	Speed and accuracy	Human rating	Modified Moses system was slightly faster and more accurate than GT at translating menu items
Wu [21]	Comparison of modified Moses system (BioMT) with GT to translate biomedical literature into multiple languages	SMT	Fluency, accuracy	Human rating and automated rating using BLEU[c]	Mixed results. Best results with translations of common European languages (German, Spanish French) with larger training corpora (100 K).
Anazawa [41]	Comparison of GT and manual translations of nursing literature abstracts	SMT	Intelligibility (quality) and perceived usefulness	Human rating	GT judged not to have sufficient quality, but perceived as useful. Longer word count and technical terms decreased accuracy.
Anazawa [42]	Comparison of GT vs. Bing Translate, Cross Language, BizLingo for nursing literature	SMT	Accuracy and intelligibility	Human rating using Proper Translation Rate	GT performed best in terms of intelligibility and PTR but quality is still not acceptable.
Dwivedi [48]	Analysis of sentence structure and accuracy of MT translation from English to Hindi for homeopathic medicine	RBT	Sentence structure and accuracy	Automated rating using BLEU	System accuracy was considered good (82%).
Mandel [44]	Comparison of traditional translation processes with and use of MT	SMT	Workflow, task analysis, subjective attitude toward MT	Qualitative interviews, workflow analysis	Bilingual staff currently use Google Translate for English to Spanish with post-editing. Express positive attitude toward MT.
Turner [14]	Comparison of GT and human post editing with manual translations	SMT	Translation workflow, cost, time, accuracy, fluency, and quality	Workflow analysis and human rating	English to Spanish using Google Translate plus post-editing is equivalent to manual translation in rating by bilingual translators.
Patil [36]	Performance of GT of common medical statements from English to 26 languages	SMT	Accuracy of translation, preservation of meaning based on back translation	Human rating accurate or not	Accuracy varied by language. In total only 57.7% of the translations were rated correct. Concluded that although useful and easy to use, GT cannot be relied on for medical communication.
Taylor [47]	Comparison of GT vs Babylon 9 for translating clinical research texts into 24 languages.	SMT	Quality	Human rating using Translation Quality Assessment (TQA) tool	Quality of GT better than Babylon however quality was still not considered acceptable. Post editing required. Large variation by languages: Somali 10% accurate, Portuguese 90% accurate.
Turner [31]	Comparison of GT with human post editing (MT + PE) vs. human translation (HT) of public health materials English to Chinese.	SMT	Time for MT post-editing, grammar, accuracy, word choice, cultural appropriateness and readability	Human rating using linguistic error analysis	GT with post editing not as good as human translation. Quality not sufficient for public health use.
Liu [51]	Comparison of hybrid Noteaid system vs GT, Bing Translator translation of medical records notes	Hybrid/SMT	Errors and preference in blind comparison	Human error rating and automated rating using BLEU	Hybrid system simplifies text first. GT performed better than hybrid system in terms of error rate.

(*continued on next page*)

Table 4 (*continued*)

Article	What was evaluated	MT Approach	Parameter Measured	Evaluation Instrument	Key Findings
Costa-Jussa [43]	Comparison of Translendium (RBT) vs UBC/Babylon 9 (SMT) to translate journalistic text and medical text from Spanish to Catalan.	RBT/SMT	Accuracy, error rate, and preference	Human rating and automated rating using BLEU and Translation Error Rate (TER)	Translation performance much better for journal text than medical text. SMT better for Spanish to Catalan as judged by human raters and BLEU. But from medical translations Catalan to Spanish human rating was higher for RBT.
Muhaxov [34]	Combines a parallel corpora system and Moses SMT software for translation of questionnaires from Chinese to Kazik and Uyghher	Hybrid	Lacks details about evaluation criteria	Human rating	Hybrid system parallel corpora machine translation 60% effective.
Chen [25]	Comparison of GT vs human translation of patient diabetes pamphlet from English to Chinese, Spanish	SMT	Fluency, accuracy, meaning and severity	Human rating	GT did adequate for simple sentences but poorly for higher complexity. English to Chinese human translation judged better.
SPEECH SYSTEMS					
Bouillon [39]	Performance of Medical speech translation system (MedSLT)	RBT	Translation completion	Human rating	36/42 questions judged to have been translated completely (bidirectionally)
Bouillon [40]	Compares MedSLT to commercial systems of diagnostic questions.	RBT	Quality	Human rating using three-point scale	Sentences recognized well were translated into well into Arabic. Sentences containing technical terms were poorly translated.
Ehsani [37]	Performance of modular speech translation system (S-MINDS)	RBT and SMT	Accuracy and satisfaction	Human rating using Laws' four-point scale	High accuracy and satisfaction but whenever off script it did not work well.
Starlander [27]	Compares evaluation criteria using MedSLT	RBT	Accuracy, sentence error rate (SER) Semantic Error Rate (SemR), task completion, usability	Human rating and automatic rating using BLEU	SER ~40% but reviewers preferred unrestricted over less error-prone restricted. Error rate less as rated by human and automatic.
Ozaki [24]	Parallel text system (voice translation and response to yes/no and free-text answers)	SMT	Scenario-based task completion and accuracy	Human rating using 5-point scale	Evaluation tasks greater efficiency with fixed questions yes/no answers.
Soller [38]	Comparison of concept-based speech system (S-MINDS) vs. GT, Dragon, Jibbigo	SMT	Accuracy, fluency, usability and performance (in noisy vs. quiet settings)	Human rating using Laws' 4-point accuracy scale	Concept-based system performed better for speech-to-speech translation.
Seligman [32]	Performance of Converser in clinical setting	RBT	Usability	Human rating	Problems but positive attitude toward system.
Shin [35]	Performance of speech-to-speech robot translation English to Korean	RBT	Scenario-based task completion rate	Human rating and automated rating using METOR	Speech recognition is better with native speakers, translation was accurate for limited number of sentences.

[a] GPHIN-Global Public Health Information Network.
[b] METOR- Metric for Evaluation of Translation with Explicit Ordering.
[c] BLEU- Bilingual Evaluation Understudy score.

benchmarks for quality across studies was problematic. For example, the GPHIN was deemed "effective" by catching 56% of disease outbreaks with its combination of "best of breed" methods, which were not described in any detail [50]. The reported quality of MT systems varied depending on the MT method used (if it was described in any detail at all), language, source, and method of evaluation. However, making comparisons across these studies is complicated by the incommensurability of different languages (e.g., English to Japanese, Chinese, Spanish, French, Hungarian, Polish, Turkish, German), source documents (e.g., health promotion materials, journal abstracts, electronic health record notes, clinic conversations), and evaluation measurements (human vs automatic) involved (See Table 4 below).

4.5.1. Evaluation of text translations

Evaluations of text translation systems generally involved a scenario or task-based human evaluation of translation quality (accuracy, fluency) and/or system usability. Translation quality was most frequently measured with an accuracy scale. However, scales varied from good-bad, to application of Laws' [53] five-point scale (good, fair, poor, mistranslated, not translated) [37,38]. Most evaluations of text systems were performed on unidirectional systems involving a single translation pair. In addition, automated error analysis (BLEU score) was performed in conjunction with human translation in three studies [21,43,51] and alone in one [48].

Many of the text evaluations involved Google Translate either as the primary translation tool or as a comparison for RBT systems or other SMT systems. One of these studies, investigating 26 languages, found that the system was more accurate with Western European than with African or Asian languages [36]. Zeng-Treitler et al. employed Babel Fish to translate medical record sentences from English to Spanish, Chinese, Russian, and Korean; 76–92% of the translations were found to be incomprehensible and 77–89% were incorrect. The investigators concluded that MT was not adequate for use in medical domains [49]. On the other hand, a custom RBT English to Hindi system, used in homeopathic medicine, reported an accuracy rate of 82.3% [48]. A comparison of RBT versus SMT translation of lay and medical materials, from Spanish to Catalan, found the SMT system provided more accuracy of translation of lay and medical text; however, due to their perceived randomness, errors made by the SMT systems were reported to be more irritating than those made by the RBT system [43]. In addition, systems performed better with simplified sentence structure over complex and with technical terms pre-edited for lay terms prior to translation.

4.5.2. Quality concerns

No matter what the language or form of MT, all studies indicated that MT error rates were currently unacceptable for actual deployment in health settings. Two evaluations examined the severity of problems stemming from the few mistranslations produced; they found no significant difference between the severity of errors made by human and Google Translate for Spanish [25], but more frequent and dangerous mistranslations in Chinese [25,31]. Interestingly, one study comparing two pilot speech-to-speech translation systems showed that users preferred a system that works "well enough" to avoid dangerous interactions without constraining their ability to express themselves [30]. Higher reading level source documents were correlated with poorer MT accuracy [25].

Although many evaluations confirmed that use of MT could improve efficiency and costs, concerns about inaccuracy underscored the need to use MT as a starting point only [31,34,36,41,42,44,45,47,50]. In general, these articles identified a need to improve translation results, either through post-editing, where experts make corrections to the translations, or by enhancing MT system training for users to enhance quality, particularly with regard to specialized domain vocabulary. About a quarter of the articles reported efforts to improve translation quality either through sentence simplification pre-translation [51], using fixed phrases [24], adding images [33], or performing post-translation editing [45,14,31,44]. Anazawa et al. examined the comparative performance of several online MT systems [42], and nurses' perceptions of their effectiveness and usability [41,42]. They found MT usefulness limited without greater English language training for the nurses to interpret the poor MT output [41,42]. Several other articles highlighted accuracy, usability, and other evaluation scores that were strong but not perfect, with most recommending better training and evaluation to improve MT performance [21,25,26,32,33,35,37,39,40,43,46,48,49].

In speech-to-speech translations, accuracy must be balanced with the ability to converse. For example, in the MedSLT project, Starlander and Estrella tested a version of the system that was constrained to yes-no and short elliptical responses against a second version that allowed users more freeform responses [30]. This study found that users preferred the less accurate version of their system; imperfect translations that still avoid dangerous interactions were favored over more accurate but restricted interactions. Seligman et al. described a speech system (The Converser) which allows the user to correct the original speech input and save the preferred translations for future system training and use [32].

4.6. What are the areas for future work?

The areas for future work highlighted by the articles in this review helped to illuminate what gaps need to be addressed from the technology developers' perspectives. Most studies did not involve the actual implementation of systems within a clinical context, but rather tested and evaluated systems with selected groups of users or through simulations, and many of the recommendations for future work are derived from these tests.

Three studies concluded that existing MT methods are not suitable for clinical communication, due to limitations in quality and considering the potential implications of incorrect medical translation [36,47,49]. Despite these limitations in clinical practice, one study offered suggestions for where MT could currently be useful, such as improving bedside manners by allowing providers to ask individuals in the hospital about their basic needs (such as food requests or room-related requests) and have non-critical conversations [26]. Only one study examined workflow in the context of clinical communication and evaluated the use of MT during the clinic's intake process to better understand patient symptoms or complaints [33]. Several other researchers worked to develop improved systems for communicating between patients and providers in clinical practice, and recommended follow-up studies on their speech-to-speech [24,38–40] and text-based systems [33,34] for validation of their findings.

In settings focused more on health education, such as public health, discussions and recommendations more frequently examined workflow. For example, one qualitative study (in a public health department where health-education materials were being translated) outlined the current human translation workflow at the agency [44]. The process of human translation was time and resource intensive for employees, but considered critical to ensure translation quality. Considering its low cost, some employees interviewed in the study thought about incorporating MT into the workflow as a starting point for translation, with additional human post-editing [44]. Other studies that incorporated a workflow perspective in the public health setting similarly recommended MT as a starting point, to be followed by post-editing, in order to save both time and costs in the translation process [45,14]. In a nursing setting where staff were seeking continuing education [41], researchers pointed to the need for more training in how to effectively use MT tools, in order to facilitate their adoption. Two studies also emphasized the need for a greater focus on design based on user needs rather than on technology-driven solutions [41,44].

The studies offered numerous areas for future work to improve the research around MT quality. Suggestions included developing uniform quality metrics for MT evaluation [30] or domain-specific evaluation

metrics [51]; further evaluating MT system errors and how they impact human judgement of translation quality [43]; and providing more description and critical assessment of translation methods in peer reviewed literature related to translation [47]. Several authors also offered suggestions for improving MT technology. For example, some suggested focusing on narrower domains to improve the vocabulary and quality of MT within those domains [42,48,49]. Others pointed to a need for better source language training, mentioning that MT algorithms produced better quality translation output for languages more frequently used in computing, such as Western European languages [36], and offered suggestions such as expanding training corpora and having parallel corpora across languages to compare quality [21]. Others reported on ways to improve the accuracy of MT, such as a universal MT code system [25], adding 'n-gram' steps to systems [46], on-the-fly context disambiguation and terminology management [45], and semantic role labeling or abstract meaning representation [51].

5. Discussion

Significant strides were made in machine translation technology over the decade from 2006 to 2016, including both advances in statistical modeling as well as increased amounts of data available for speech and text translation. As a result, machine translation performance improved dramatically for languages and topic areas where sufficient corpora were available for training. Although there was only one publication describing an actual deployment of machine translation in a health setting during this period, the advances in machine translation technology led to the development and pilot testing of a variety of text and speech translation systems for potential use in clinical and public health domains.

MT is currently being developed in the health communications field for two main purposes: to improve patient-provider and patient-staff communication in multilingual clinical settings, and to increase access to health education resources in minority languages. Our second main research question within these two application domains regarding which approaches show strongest evidence of promise for adoption, remains unclear due to the lack of shared evaluation criteria. Only two systems were tested in a field setting, and even where pilot systems were developed and tested, numbers of participants were limited and evaluation methods across papers were inconsistent, making comparisons difficult. We need better criteria for how to evaluate efficacy and a shared understanding of metrics for accuracy, speed, and other quality measures. There is room to work with targeted users to develop evaluation frameworks and metrics well suited for their needs in real-life settings. The Translation Automation User Society (TAUS), for example, has highlighted evaluation as a significant barrier to the efficient adoption of MT in practice, proposing its Dynamic Quality Evaluation Framework as a remedy for the lack of common best practices, benchmarking data, and learning exchange between academia and industry in MT quality evaluation. Its associated online toolkit and benchmarking database are available to members [54].

Given the necessity of accurate translation in health care settings, even systems that showed high translation accuracy according to their selected evaluation metric (e.g. BLEU score) might not be good enough to encourage adoption. As several studies noted, the most promising text translation solution at this time is to use MT as an initial, time- or money-saving step, with subsequent correction or verification of accuracy by a human translator with domain expertise.

In highlighting the need for better MT quality across a variety of languages, studies pointed to multiple avenues for improving output. To ensure quality for SMT, a sufficient amount of domain-relevant parallel training data needs to be available [21]. An approach highlighted by Pozo et al. is to increase the domain-specific training, so the underlying MT engine can better handle specialized vocabularies and grammars [46]. In the meantime, graphics and images (such as pictures of the body) and other multimedia add-ons could help clarify translations that are not yet adequate in quality [33,46].

Practitioners considering adopting speech translation tools have additional considerations besides translation quality. The ability to input speech can vary dramatically based on individual differences in speech characteristics, such as pitch and accent [35]. Attention must also be paid to selecting a quality close range microphone and other appropriate hardware [39], since clinical environments can be noisy and have multiple speakers within range of a microphone. In addition, protocols for securing any personal health information should be in place before introducing a new translation device to a clinical setting.

The most promising of the speech technologies, and the one that appears closest to actual clinical deployment, was the S-MINDS concept translator, which uses a combination of rule-based and SMT methodology [38]. S-MINDS was rated high for accuracy and patient satisfaction. Unfortunately, interpretation of these results is difficult because the article did not provide sufficient detail regarding the extent to which the system was incorporated into the clinical setting or how it was evaluated.

In general, publications involving the same system but published serially were often difficult to interpret alone due to incomplete descriptions of details from prior studies. Standards for conducting and reporting evaluation of information systems, such as the proposed STARE-HI standards for evaluation studies in health informatics [55,56] would greatly facilitate their interpretation to readers, who may not be familiar with earlier results.

Overall, there is an immense need for real-world deployments and validation studies. The studies in this review skewed toward pilots and explorations, with little work done in real-life settings, over extended periods of time, using validated evaluation methods. The Pan American Health Organization Machine Translation System (PAHOMTS) has been in use since 1980, producing fast and inexpensive health translations, and it has more than 100,000 English-Spanish-Portuguese health dictionary entries in each language [57]. However, since the PAHOMTS-related publication fell outside of our study timeframe, it was excluded from our review. In the absence of full deployment among the selected 27 articles except [50], scenarios and task-based experiments were the favored methods of evaluation.

The articles in this review pointed to improving the underlying MT technology, increasing MT training, and better incorporating human judgement and correction as the main directions for further work. If MT is to be adopted widely in service of health communications, there is a need for more foundational empirical research to support these systems, given the potential for harm if there are errors in translation of health information. In addition to longitudinal deployment and validation studies, the field could benefit from constructing a set of benchmark tasks, for example, collecting medical documents of different types (e.g., medical records, discharge summaries, and consult notes). These tasks could be used to test MT systems and to evaluate their performances with shared criteria (e.g. BLEU score), making it easier to interpret the outcome data. This could be complemented by work that creates standard health communications use cases for the various settings inside and outside the clinic, and then evaluates MT systems against existing practices regarding time efficiency, cost savings, and patient satisfaction scores.

The studies cited above used MT approaches or systems that predate the more recent neural MT paradigm. We are not aware of any study testing state-of-the-art neural MT in healthcare settings. In a technical implementation paper, Wolk and Marasek used multiple automatic evaluation metrics to compare English-Polish SMT versus NMT models using a training corpora comprising medical product documents, finding the SMT models performed better and required only one day to perform the computations compared to 4–5 for the NMT systems [58]; however, the ongoing training and maintenance for NMT is lower, which could offset other performance concerns. Despite the overall improvement in performance that can be expected from this approach in the longer term, we expect that the problems of domain

specialization and the need for utmost accuracy in healthcare settings will continue to require intensive engineering efforts. Ongoing work in domain adaptation techniques [59,60] should help better tune models to health communications domains. In addition, unless better practices for adoption and evaluation are developed, this improvement may go unnoticed in the health community.

6. Limitations

The primary strength of this review is also the source of its greatest limitation. Our interdisciplinary systematic search approach covered engineering, nursing, public health, clinical health, and linguistics databases to gather as much relevant work as possible. It is this interdisciplinary nature of MT research in domain-specific settings that makes it difficult to draw conclusions about which approaches are most promising, since each discipline and domain comes with its own empirical traditions, preferred methods, and evaluation criteria. We acknowledge that a quality rating for each article would be helpful for indicating the strength of evidence for the claims made. However, because the included articles spanned qualitative, quantitative and mixed methods approaches, no single quality-rating framework (e.g. the STARE-HI for qualitative health informatics research [55,61]) was appropriate across all articles. Similarly, certain PRISMA items, e.g., the risk of bias assessment, were not included here because such items have underlying epistemological commitments that do not apply to qualitative and mixed-methods research. As with all qualitative work, the screening and full text review could have been impacted by interpretive "bias" among the researchers. The authors' own work in this area was returned by the systematic search, but we were diligent in applying the same inclusion/exclusion criteria for all articles and excluded some of this work. Furthermore, by limiting our search to English language articles, we have excluded relevant work being conducted in other languages.

7. Conclusion

In this paper, we first reported on the current state of MT technology in development to overcome linguistic barriers in health settings based on our systematic search and qualitative analysis of 27 studies. We found that current developments of MT technology primarily aim to improve in-clinic communication between patient and provider or to generate multilingual health education materials. The target users for MT applications were mostly patients who do not speak the dominant language of their medical environment. The majority of studies involved unidirectional translations from a single source language to a single target language, with English-Spanish being the most common language pair. Although the reviewed studies showed the potential use of MT systems in clinical settings has been an area of increasing interest over the last decade, actual incorporation of MT into the clinical environment remains very limited. Continued concern about the accuracy and fluency of MT in the health domain, where mistakes and misunderstandings could have dramatic consequences, still hinders the usefulness of the technologies in real clinical settings.

The underlying MT approaches were roughly split between SMT and RBT. While the majority of systems targeted text translation, several applications were designed for real-time speech interaction. Only two text systems were based exclusively on RBT [48,49], with the remainder being hybrid systems or not clearly described; the speech systems more frequently utilized a RBT [35,38–40] or hybrid [24,37] approach.

Regarding which approaches show evidence of promise for adoption in health settings, all but two of the studies performed an evaluation involving a MT system, but the nature and strength of the evaluation findings varied dramatically. Furthermore, only two were evaluated in their respective field settings [32,50], but even those lacked performance benchmarks. The reported quality of MT systems varied depending on the MT method used (if described), the different languages and source documents, and the variety of evaluation methods and metrics (human vs automatic) involved.

However, there was consensus that MT technology alone produces inadequate quality translations for healthcare settings; of all the systems described, about half reported on the use of human analysis and correction to improve the final output of translations. Suggestions for future work included developing uniform quality metrics for MT evaluation [30] or domain-specific evaluation metrics [51], further evaluating MT system errors and how they impact human judgement of translation quality [43], and providing more description and critical assessment of translation methods in peer reviewed literature related to translation [47]. Several authors also offered suggestions for improving the underlying MT performance. There was evidence that using MT as a starting point in conjunction with human correction is a fruitful strategy for preserving accuracy while benefitting from MT's cost and speed advantages.

The interdisciplinary nature of this work, the variety of systems described, and the range of evaluation methods employed make it difficult to draw clear conclusions about which MT approaches are most promising. In healthcare, similar to what has been described in other domains where large corpora are available, SMT appears to be more accurate than RBT systems. Improved accuracy and broader implementation, coupled with standardization of evaluation frameworks are needed to harness successfully the potential of MT in improving health communications in multilingual settings.

Conflict of interest

Dr. Kirchhoff has served on the Advisory Board of the European HiML (Health in my Language) project and is currently under contract with Google.

Funding

This study was funded by grant #1R01LM010811 from the National Library of Medicine (NLM). Its content is the sole responsibility of the authors and does not necessarily represent the views of the NLM. The authors would like to thank Beryl Schulman for her review of this manuscript.

References

[1] C. Pandya, M. McHugh, J. Batalova, Limited English Proficient Individuals in the United States: Number, Share, Growth, and Linguistic Diversity. LEP Data Brief, Migration Policy Institute, 2011.
[2] C. Ryan, Language use in the United States: 2011, American community survey reports, 2013, p. 2.
[3] N.A. Ponce, R.D. Hays, W.E. Cunningham, Linguistic disparities in health care access and health status among older adults, J. Gen. Intern. Med. 21 (7) (2006) 786–791, https://doi.org/10.1111/j.1525-1497.2006.00491.x.
[4] T.L. Sentell, J.Y. Tsoh, T. Davis, et al., Low health literacy and cancer screening among Chinese Americans in California: a cross sectional analysis, BMJ Open 5 (2015) e006104, https://doi.org/10.1136/bmjopen-2014-006104.
[5] N. Peña-Purcell, Hispanics' use of Internet health information: an exploratory study, J. Med. Library Assoc. 96 (2) (2008) 101.
[6] A.J. Cardelle, E.G. Rodriguez, The quality of Spanish health information websites: an emerging disparity, J. Prevent. Intervent. Commun. 29 (1–2) (2005) 85–102.
[7] E.M. Cheng, A. Chen, W. Cunningham, Primary language and receipt of recommended health care among Hispanics in the United States, J. Gen. Intern. Med. 22 (2) (2007) 283–288.
[8] L. Shi, L.A. Lebrun, J. Tsai, The influence of English proficiency on access to care, Ethnicity Health 14 (6) (2009) 625–642.
[9] Translating rights into access, Language access and the affordable care act, Am. JL Med 38 (2012) 348.
[10] Civil Rights Act of 1964 § 7, 42 U.S.C. § 2000d et seq (1964).
[11] Office of Civil Rights, Guidance to Federal Financial Assistance Recipients Regarding Title VI Prohibition against National Origin Discrimination Affecting Limited English Proficient Persons, 2002, Available from: <http://www.lep.gov/guidance/guidance_index.html. Retrieved 11-27-16>.
[12] Office of Minority Health, National Standards for Culturally and Linguistically Appropriate Services (CLAS) in Health and Health Care, Available from: <https://

www.thinkculturalhealth.hhs.gov/pdfs/EnhancedNationalCLASStandards.pdf> Retrieved 11-27-16.
[13] Patient Protection and Affordable Care Act, 42 U.S.C. § 18001 et seq. (2010).
[14] A.M. Turner, M. Bergman, M. Brownstein, et al., A comparison of human and machine translation of health promotion materials for public health practice: time, costs, and quality, J. Public Health Manage. Practice 20 (5) (2014) 523–529.
[15] N. Lunt, R. Smith, M. Exworthy, Medical Tourism: Treatments, Markets and Health System Implications: A Scoping Review, Organisation for Economic Co-operation and Development, Paris, 2011.
[16] C. Muller, Machine translation post-editing holds the key to MT success, Available from: <http://www.csoftintl.com/knowledge_vault/machine_translation_post_editing_holds_the_key_to_mt_success> Retrieved 11-27-16.
[17] N. Cancedda, M. Dymetman, G. Foster, et al., A statistical machine translation primer, in: C. Goutte, N. Cancedda, M. Dymetman (Eds.), Learning Machine Translation, The MIT Press, Cambridge, Mass, 2009, pp. 1–37.
[18] P. Koehn, F.J. Och, D. Marcu, Statistical phrase-based translation, Proceedings of the Meeting of the North American Association for Computational Linguistics (NAACL), Association for Computational Linguistics, Edmonton, Canada, 2003, pp. 48–54.
[19] E. Oladosu, A. Esan, I. Edayanju, et al., Approaches to machine translation: a review, J. Eng. Technol. 1 (1) (2016) 120–126.
[20] R.J. Weiss, J. Chorowski, N. Jaitly, et al., Sequence-to-sequence models can directly transcribe foreign speech, CoRR, 2017. abs/1703.08581.
[21] C. Wu, F. Xia, L. Deleger, et al., Statistical machine translation for biomedical text: are we there yet, AMIA Annual Symposium Proceedings, vol. 2011, American Medical Informatics Association, 2011, p. 290.
[22] Y. Wu, M. Schuster, Z. Chen, et al., Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, arXiv:1609.08144.
[23] D. Moher, A. Liberati, J. Tetzlaff, et al., Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, Ann. Internal Med. 151 (4) (2009) 264–269.
[24] S. Ozaki, T. Matsunobe, T. Yoshino, et al., Design of a face-to-face multilingual communication system for a handheld device in the medical field, International Conference on Human-Computer Interaction, Springer, Berlin, 2011, pp. 378–386.
[25] X. Chen, S. Acosta, A.E. Barry, Evaluating the accuracy of Google translate for diabetes education material, JMIR Diabetes 1 (1) (2016) e3, , https://doi.org/10.2196/diabetes.5848.
[26] R.R. Khanna, L.S. Karliner, M. Eck, et al., Performance of an online translation tool when applied to patient educational material, J. Hospital Med. 6 (9) (2011) 519–525.
[27] M. Starlander, P. Bouillon, G. Flores, M. Rayner, N. Tsourakis, Comparing two different bidirectional versions of the limited domain medical spoken language translator MedSLT, in: Proceeding of the 12th Annual Conference of the European Association for Machine Translation (EAMT), pp. 174–179.
[28] A.M. Turner, L. Desai, K. Dew, N. Martin, K. Kirchoff, Machine assisted translation of health materials to chinese: an initial evaluation, MEDINFO (2015) 979.
[29] M. Seligman, M. Dillinger, Real-time multi-media translation for healthcare: a usability study, in: Proceedings of the 13th Machine Translation Summit, 19–23 September 2011, Xiamen, China.
[30] M. Starlander, P. Estrella, Relating recognition and translation quality with usability of two different versions of MedSLT, in: MT Summit XII: proceedings of the twelfth Machine Translation Summit 2009 August 26–30, Ottawa, Ontario, Canada, pp. 324–331.
[31] A.M. Turner, K.N. Dew, L. Desai, et al., Machine translation of public health materials from english to chinese: a feasibility study, JMIR Public Health Surv. 1 (2) (2015).
[32] M. Seligman, M. Dillinger, Evaluation and revision of a speech translation system for healthcare, in: Proceedings of the 12th International Workshop on Spoken Language Translation, 2015 Dec 3–4, Da Nang, Vietnam.
[33] T. Fukushima, T. Yoshino, A. Shigeno, Development of multilingual interview-sheet composition system to support multilingual communication in medical field, International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Springer, Berlin, 2011, pp. 31–40.
[34] H. Muhaxov, Z. Lou, S. Tayila, et al., Multiple-language translation system focusing on long-distance medical and outpatient services, in: 2016 IEEE Second International Conference on Multimedia Big Data, IEEE, pp. 471–475.
[35] S. Shin, E.T. Matson, J. Park, et al., Speech-to-speech translation humanoid robot in doctor's office, Automation, Robotics and Applications (ICARA), 2015 6th International Conference, IEEE, 2015, pp. 484–489.
[36] S. Patil, P. Davies, Use of Google Translate in medical communication: evaluation of accuracy, BMJ 15 (349) (2014) g7392.
[37] F. Ehsani, J. Kimzey, E. Zuber, et al., Speech to speech translation for nurse patient interaction, in: 22nd International Conference on Computational Linguistics, 2008, p. 54.
[38] R.W. Soller, P. Chan, A. Higa, Performance of a new speech translation device in translating verbal recommendations of medication action plans for patients with diabetes, J. Diab. Sci. Technol. 6 (4) (2012) 927–937.
[39] P. Bouillon, G. Flores, M. Starlander, et al., A bidirectional grammar-based medical speech translator, Proceedings of the Workshop on Grammar-Based Approaches to Spoken Language Processing, Association for Computational Linguistics, 2007, pp. 41–48.
[40] P. Bouillon, S. Halimi, M. Rayner, et al., Adapting a medical speech to speech translation system (MedSLT) to Arabic, Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Association for Computational Linguistics, 2007, pp. 41–48.
[41] R. Anazawa, H. Ishikawa, K. Takahiro, Evaluation of online machine translation by nursing users, Comput. Inform. Nurs. 31 (8) (2013) 382–387.
[42] R. Anazawa, H. Ishikawa, M.J. Park, T. Kiuchi, Online machine translation use with nursing literature: evaluation method and usability, CIN: Comput., Inform., Nurs. 31 (2) (2013) 59–65.
[43] M.R. Costa-Jussa, M. Farrus, J.B. Marino, et al., Automatic and human evaluation study of a rule-based and a statistical Catalan-Spanish machine translation systems, in: International Conference on Language Resources and Evaluation "Seventh Conference on International Language Resources and Evaluation", 2011, Valletta, Malta, pp. 1707–1711.
[44] H. Mandel, A.M. Turner, Exploring local public health workflow in the context of automated translation technologies, AMIA Annual Symposium Proceedings, vol. 2013, American Medical Informatics Association, 2013, p. 939.
[45] K. Kirchhoff, A.M. Turner, A. Axelrod, et al., Application of statistical machine translation to public health information: a feasibility study, J. Am. Med. Inform. Assoc. 18 (4) (2011) 473–478.
[46] A.P. Pozo, A.W. Haddad, M. Boutin, et al., A hand-held multimedia translation and interpretation system for diet management, 2011 IEEE International Conference on Multimedia and Expo, IEEE, 2011, pp. 1–6.
[47] R.M. Taylor, N. Crichton, B. Moult, et al., A prospective observational study of machine translation software to overcome the challenge of including ethnic diversity in healthcare research, Nurs. Open 2 (1) (2015) 14–23.
[48] S.K. Dwivedi, P.P. Sukhadeve, Comparative structure of Homoeopathy language with other medical languages in machine translation system, Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on, IEEE, 2013, pp. 775–778.
[49] Q. Zeng-Treitler, H. Kim, G. Rosemblat, et al., Can multilingual machine translation help make medical record content more comprehensible to patients? Stud. Health Technol. Inform. 160 (Pt 1) (2009) 73–77.
[50] M. Blench, Global public health intelligence network (GPHIN), Proceedings of the Conference of the American Machine Translation Association, AMTA, Waikiki, Hawai'i, 2008.
[51] W. Liu, S. Cai, B.P. Ramesh, G. Chiriboga, K. Knight, H. Yu, Translating Electronic Health Record Notes from English to Spanish: A Preliminary Study, ACL-IJCNLP 2015, 30 (2015) 134.
[52] V. Braun, V. Clarke, Using thematic analysis in psychology, Qual. Res. Psychol. 3 (2) (2006) 77–101.
[53] M.B. Laws, R. Heckscher, S. Mayo, W. Li, I. Wilson, A new method for evaluating the quality of medical interpretation, Med. Care 42 (1) (2004) 71–80.
[54] Enabling Better Translation, TAUS, <https://www.taus.net/evaluate/dqf-tools>, Published 2017, Retrieved 6-23-17.
[55] J. Talmon, E. Ammenwerth, J. Brender, et al., STARE-HI—statement on reporting of evaluation studies in health informatics, Int. J. Med. Inf. 78 (2009) 1–9, https://doi.org/10.1016/j.ijmedinf.2008.09.002.
[56] N.F. Keizer, J. Talmon, E. Ammenwerth, et al., Systematic prioritization of the STARE-HI reporting items, Methods Inf. Med. (2012) 104–110, https://doi.org/10.3414/ME10-01-0072.
[57] J. Aymerich, Using Machine Translation for fast, inexpensive, and accurate health information assimilation and dissemination: experiences at the Pan American Health Organization, in: 9th World Congress on Health Information and Libraries, 2005, Salvador-Bahia, Brazil.
[58] K. Wolk, K.P. Marasek, Translation of medical texts using neural networks, Int. J. Reliable Qual. E-Healthcare 5 (4) (2016) 51–66.
[59] P. Pecina, A. Toral, J. Genabith, Simple and effective parameter tuning for domain adaptation of statistical machine translation, in: Proceedings of COLING 2012, 2012, Mumbai, India.
[60] R. Rubino, S. Huet Lefèvre F, G. Linarès, Statistical Post-Editing of Machine Translation for Domain Adaptation, 2012, Trento, Italy.
[61] J. Brender, J. Talmon, N.F. Keizer, et al., STARE-HI – statement on reporting of evaluation studies in health informatics, Appl. Clin. Inform. 4 (2013) 331–358.