How Emotional and Contextual Annotations Involve in Sensemaking Processes of Foreign Language Social Media Posts

HAJIN LIM, Cornell University, USA
DAN COSLEY, Cornell University, USA
SUSAN R. FUSSELL, Cornell University, USA

The goal[1] of this paper is to investigate how computational tools to annotate communication can support multilingual sense-making on social media. We conducted a field study of SenseTrans, a browser extension that uses sentiment analysis and named entity extraction techniques to annotate Facebook posts with emotional and contextual information. Interviews with 18 participants who used SenseTrans in their Facebook newsfeed for two weeks suggest that the annotations often supported sensemaking by providing additional information they could use to get a quick gist of the posts or to supplement their own interpretations. Participants varied in the extent to which they were motivated to evaluate the credibility of and form mental models of how the annotations were generated, which shaped how they utilized the annotations for sensemaking. Our findings demonstrate the value of designing to support cross-lingual communication and inform design implications for intelligent tools that support communication and sensemaking.

CCS Concepts: • **Human-centered computing** → Collaborative and social computing

KEYWORDS

Sensemaking, cross-lingual sensemaking, cross-cultural sensemaking, intercultural sensemaking, multilingual communication, intercultural communication, natural language processing, social media, machine translation

ACM Reference format:

Hajin Lim, Dan Cosley and Susan R. Fussell. 2019. How Emotional and Contextual Annotations Involve in Sensemaking Processes of Foreign Language Social Media Posts. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW. Article 69 (2019), 17 pages. https://doi.org/10.1145/3359171

1 INTRODUCTION

Social media sites have become key platforms for maintaining diverse relationships and sharing information. To support these goals, many social media sites offer a real-time feed of messages from one's connections; casual browsing and sensemaking of these messages can support these relational and informational goals [29][44][45]. As many of these social media sites are global and multilingual, users often have personal connections who speak other languages, or follow accounts and pages in other languages to keep track of their interests, learn foreign languages,

This.is work is partially supported by the National Science Foundation, under grant IIS-1421929 and grant IIS-1910147.
Author's addresses: H. Lim, Information Science, Cornell University, Ithaca, NY, USA, email: hl934@cornell.edu; D. Cosley, Information Science, Cornell University, Ithaca, NY, USA; email: drc44@cornell.edu; S. R. Fussell, Information Science, Cornell University, Ithaca, NY, USA; email: sfussell@cornell.edu

2573-0142/2019/11 - 69 $15.00 Copyright is held by the owner/author(s). Publication rights licensed to ACM. https://doi.org/10.1145/3359171

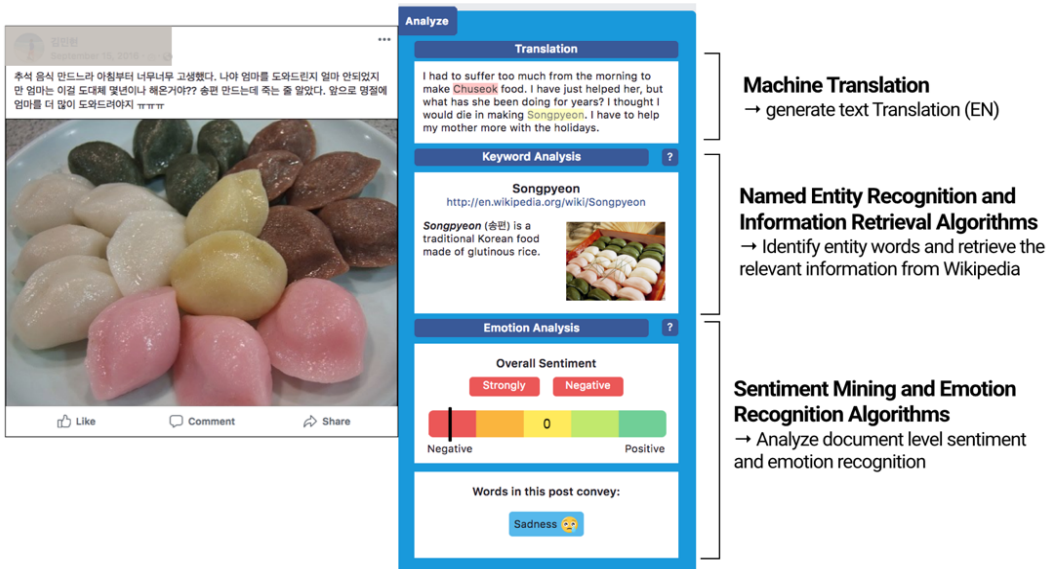

Figure 1. Components of SenseTrans and the Related NLP Algorithms. Translation (top) presents MT outputs of the text content of a post. Keyword Analysis (middle) presents additional information about key entities in the translation. Emotion Analysis (bottom) visualizes the overall sentiment of and emotions present in a post.

and get up-to-date information from other countries [19]. For example, a Facebook user who is a monolingual native English speaker might have a bilingual friend who sometimes posts in Japanese, or might have followed the Twitter account of a sports player who speaks Portuguese. In this paper, we will refer to status updates in languages cannot read or speak fluently without using translations as "foreign language posts".

As it is increasingly common to encounter foreign language posts in one's feed [1][9], social media sites including Facebook, Instagram, and Twitter have incorporated machine translation (MT) features (e.g., [46]) into their interfaces to help users to understand and interact with posts in foreign languages. While the overall quality of MT continues to improve, recent studies [24] show that social media users still have trouble making sense of foreign language posts because MT quality is inconsistent, particularly given the colloquial nature of much social media content [4], [15]. Even when MT produces reasonable translations, these translations often do not help readers understand the cultural and contextual meanings around foreign language posts [24][35]. These difficulties and information gaps, along with the cognitive load of working with a foreign language, often make people reluctant to pay attention to foreign language posts [14][25]. This in turn can have negative effects on their relationships and understanding of other cultures [3] [22].

To better support sensemaking of foreign language posts, Lim and colleagues developed SenseTrans (Figure 1) [23]. SenseTrans uses sentiment analysis and entity recognition algorithms from Natural Language Processing (NLP) to annotate posts with emotional and contextual information. The emotional analysis provides information about the overall tone of posts while the keyword analysis identifies proper nouns and links them to information from Wikipedia to help people learn about culturally specific references. In lab evaluations using mock Facebook profiles, participants using SenseTrans reported improved comprehension of, and willingness to engage with, foreign language posts, with no increase in cognitive load versus participants using MT alone [23]. These results provide preliminary support for the idea of computational annotation to support cross-lingual sense-making on social media.

However, it is still an open question how useful annotations would be for sensemaking of foreign language posts on people's own Facebook feeds. In this paper, we address this question through a study of 18 participants who used SenseTrans on their own Facebook news feeds for two weeks. Based on system usage log and interview data, we concluded that people often found the annotations useful, supporting sensemaking by helping people form and confirm their understanding of a post and quickly get the gist of posts. This in turn increased their self-efficacy, cultural learning, and feelings of connectedness with friends who posted in other languages. Participants' assessments of the value of SenseTrans, and how they ended up using it, were shaped by their motivation and ability to evaluate the credibility of the algorithmic annotations and form mental models of how these algorithms worked. Our findings both demonstrate the potential of tools to support cross-lingual sensemaking and inform the design of future tools to such sensemaking.

2 BACKGROUND

In this section, we discuss the challenges involved in supporting cross-lingual sensemaking of social media posts, along with existing design approaches to supporting cross-lingual communication. We then describe how SenseTrans is designed to provide annotations to help address those challenges of making sense of foreign language posts on social media.

2.1 Making Sense of Foreign Language Social Media Posts

Social media sites support multilingual communication among diverse users who have different linguistic and cultural backgrounds [1][9]. Having access to this linguistic and cultural diversity offers a number of potential benefits such as improving cultural understanding and intercultural communicative competency [5][39]. Recognizing this, social media sites are increasingly offering MT features to help users to comprehend posts in other languages (e.g., [46]).

However, making sense of linguistically and culturally different information requires complicated sensemaking processes that MT outputs only partly support. Language differences pose not only the immediate barrier of understanding the other language, but the more general barrier of cultural and contextual differences [7][30][33] because languages are often associated with specific countries or cultures. By themselves, machine translations help address the first issue but not the second, which has been found to be a key problem for people when making sense of foreign language posts in social media [24].

The context of social media also poses challenges for cross-lingual sensemaking. For instance, MT systems struggle with the short, informal text that characterizes many social media posts. This tends to translate into increased errors and unstable quality of MT outputs that reduce people's ability to use MT for cross-lingual sensemaking in social media [24]. Together, these issues have been shown to cause people to be reluctant to pay attention to posts in other languages [25], which in turn can reduce their ability to interact with people and gain information across linguistic boundaries on social media [3][22].

2.2 Annotation and Cueing to Support Cross-lingual Communication

One design approach to supporting communication in the face of translation issues is to provide additional information, or annotations, as cues to supplement translation outputs. Researchers have used a variety of approaches to generate such annotations, including highlighting keywords [11], showing multiple MT outputs at a time [12], adding semantically related images [38], and providing back-translation [34]. These techniques can provide valuable information for diagnosing and recovering from communication errors caused by translation errors; lab experiments suggest these approaches can improve message clarity [11][12] and collaboration outcomes [38].

Many of the annotations described above focus on helping people comprehend the literal meaning of others' messages clearly by recognizing, repairing, and recovering from translation errors in synchronous chat contexts. Further, these annotations could focus on providing supplemental information to help mitigate cultural and contextual barriers. This idea is inspired by Shigenobu and colleagues' AnnoChat [35], a multilingual chat system that allowed people to manually add and share annotations to messages, as well as Yuizono and colleagues' study of providing cultural knowledge support for topics arising in chat conversations between multilingual participants [43]. Lab studies of these systems suggest that annotating chats with information to provide cultural context has potential value, although a key challenge was automatically creating such annotations to reduce human effort.

2.3 SenseTrans: A Browser Extension for Sensemaking of Foreign Language Posts

Bringing the idea of annotation to support cross-lingual communication together with the need to support cross-cultural sensemaking in social media, Lim and colleagues developed a browser extension called SenseTrans [23] that aimed to address the cultural and contextual challenges in sensemaking of foreign language posts described above and in previous work [24]. To do this, SenseTrans provides two types of additional information regarding the text content of a post: *contextual knowledge* of the often culturally-specific entities referenced in a post, and *emotional knowledge* of the sentiment or feelings expressed in a post. In extracting such information, SenseTrans makes use of NLP algorithms such as named entity extraction and sentiment analysis techniques.

The interface of SenseTrans consists of three sections (Figure 1). The first, labeled 'Translation' in Figure 1, presents a machine translation results from the Google Translate API. The second, labeled 'Keyword Analysis', provides annotations about named entities extracted from a post such as landmarks, historic sites, holidays, politicians, and celebrities that might help increase readers' awareness of cultural context [35]. The third, labeled 'Emotion Analysis', provides annotations about the overall sentiment and presence of major emotions in a post, to help readers better understand the personal context of the poster. Although the SenseTrans can analyze posts in any language posts, all information presented in SenseTrans was in English.

To validate the utility and accuracy of the annotations in SenseTrans, Lim and colleagues conducted systematic summative evaluations of outputs from SenseTrans on 241 non-English social media posts written in 42 different languages provided by bilingual Amazon Mechanical Turk users. The result showed acceptable performance: the system correctly detected the sentiment (as indicated by the original poster) of 74% of status updates and keyword analysis captured 84% of the words that were culturally specific.

As described in the introduction, the initial lab study was promising: participants who used SenseTrans to browse Facebook profiles in foreign languages reported higher perceived comprehension of and willingness to interact with foreign language posts than users in the MT only condition.

However, the initial lab study had several limitations. First, it used mock Facebook profiles in a one-shot study, trading experimental control for ecological validity; thus, it's unclear how valuable SenseTrans would be for sensemaking of foreign language posts on people's own Facebook feeds. Further, because the effects of using SenseTrans were measured based on self-report Likert scales without follow-up interviews, it requires further investigation of both how users perceived such annotations and used them in their sensemaking, and what factors affected those perceptions and uses.

2.4 Research Questions

Together, prior work suggests that by itself MT is not sufficient to support effective cross-lingual sensemaking, and that annotating MT with additional cues is a potentially effective approach for supporting it. However, most work in this space has focused on synchronous chat; sensemaking around social media feeds is different in terms of people's sensemaking goals and needs, as well as the amount of effort people are willing to put in, leading to questions about how people would use such annotations in sensemaking of foreign language posts:

RQ1. How do people use the annotations in sensemaking?

Further, a number of factors might affect people's perception and understanding of these annotations. Unlike MT, which has been integrated into many web applications [17] and which people have some familiarity with and expectations around (e.g., [2][20]), other kinds of annotation are much less common. This raises questions of how people came to build their perception and understanding of new types of annotation:

RQ2. What factors shape their perceptions of the usefulness of the annotations?

Finally, we ask how the those perceptions affected the value of using these annotations in sensemaking. Do they help people overcome the barriers described earlier and achieve the potential relational and informational benefits of being exposed to different cultures and languages online?

RQ3. What are the overall outcomes of using the annotations in a real context?

3 RESEARCH METHOD

To address the research questions, we conducted a two-week in the wild deployment of SenseTrans where people used it with their own Facebook newsfeeds. The study consisted of three phases: (1) an initial set-up meeting; (2) the two-week system deployment and data logging period; and (3) an in-depth post-deployment interview.

3.1 Study Procedure

3.1.1 Phase 1: Initial Set-up

Participants visited a lab space on campus with their laptop. We first explained the overall procedure, including the types of information we would collect during the deployment phase. After participants consented, we installed the SenseTrans Chrome extension on their laptop. We explained each component of the interface, though we did not explain the underlying technical implementations, and let them explore the interface to become familiar with the features. Participants then completed a survey about their usual Facebook browsing habits. We also conducted brief interviews to understand 1) what they usually did when they encounter foreign language posts and 2) how they perceived existing MT features in making sense of the posts on social media. Finally, we set a time for the post-deployment interview two weeks later.

3.1.2 Phase 2: System Deployment and Data Logging

We asked participants to explore SenseTrans and use it to browse their Facebook newsfeed during those two weeks. For people who had relatively few foreign language posts in their newsfeed, we asked them to follow some foreign news or celebrity pages they were interested in, to increase their exposure to both those posts and the tool. Other than that, we did not give any specific instructions or tasks beyond asking them to use it as much as reasonable so they would have meaningful experiences with it.

During the deployment, we logged selected aspects of posts for which users accessed SenseTrans, including the poster's name, post time and post language, and timestamps when they

opened the keyword and emotion analysis features. We made sure not to store any of the actual contents of the post to protect the privacy of participants and their Facebook connections. These data provided an overview of usage patterns and a way to help cue participants' reflection during the post-deployment interviews.

3.1.3 Phase 3: Post Deployment In-person Interview

After the deployment, participants returned to the lab with their laptop. We first presented them with their logged data and gave them an opportunity to review it and delete any data they did not wish to share. We then conducted an in-depth interview. The first part of the interview was about their overall impression of and experience with SenseTrans. Next, we asked how they used, evaluated, and felt about each component of the system. Then, we used the log data (particularly poster names and timestamps) to help them recount interactions around specific posts. For posts participants were willing to show us, we made structured notes about features of the post that people remarked on, without taking a screenshot or collecting personally identifiable information. The interviews lasted 20–40 minutes and were audio-recorded. Before participants left, we uninstalled SenseTrans from their laptop.

3.2 Study Participants

To recruit participants, we posted the study information to an online participant recruitment platform at a large U.S. university. In the advertisements, we mentioned that we were looking for active Facebook users who were interested in trying out a new machine translation interface for Facebook. We did not impose restrictions regarding people's native language or the number of foreign connections on their Facebook networks, as we hoped to recruit people with a diverse range of types and amounts of foreign language content in their feeds. Participants were compensated with experimental participation credits that could be used to satisfy course requirements.

In all, we recruited 23 active Facebook users. Because three of them did not complete the study and the logging did not work for two, we ended up with 18 participants. Their mean age was 20.6 years (SD = 2.3) and eight of them were male. All participants were fluent in English, though seven had other native languages (5 Chinese, 1 Korean, 1 Russian). All study procedures were conducted in English.

Most participants browsed their Facebook newsfeeds at least once a day, and had on average 771 friends (Min=153, Max=1500). They estimated that around 10 percent of the status updates on their newsfeeds were written in a foreign language. About 60% (11/18) reported they occasionally tried to make sense of foreign language posts, mostly when posts had appealing visual images or were from close friends. The rest of them (7/18) said they usually skipped the posts in foreign languages.

All participants were aware of the MT feature on Facebook. Except one participant, they had used it at least occasionally to translate friends' Facebook posts, so they had some familiarity with and expectations of MT. In the pre-deployment interview, more than half of the participants (10/18) reported that they did not feel that Facebook's MT feature was very helpful in understanding foreign language posts, describing MT results as often incorrect or incomprehensible. Six of them felt it was moderately helpful as it usually gave them a broad understanding of the posts. The other two participants highly appreciated the usefulness of MT on Facebook in general.

In contrast to MT, only four participants had heard of the sentiment analysis or named entity extraction algorithms used in SenseTrans. Those four had some level of knowledge of these techniques from related courses but they had not made personal use of them; this means that on balance, participants approached these features with a clean slate.

3.3 Data Analysis

We fully transcribed data from pre and post deployment interviews and three researchers read the transcriptions multiple times. We then open-coded the interview transcripts line by line, identifying phenomena of interest and assigning tentative labels as initial codes [28] using Atlas.ti [47]. Examples of codes at this stage were "using annotations for skimming", "wondering the meaning of 'keyword'", "relying on annotations more when the MT quality is bad", and "comparing my own interpretation of a post with the annotations".

Next, we iteratively created common themes by identifying frequently occurring codes, which enabled us to develop higher-level descriptions such as "different use cases of emotion analysis", "assessing credibility of embedded algorithms", and "uses of annotations in connection with varying MT qualities". Following that, we re-sorted the themes based on the high-level categories guided by our research questions (i.e., annotation usage patterns, factors shaping the usage, and overall experiences and outcomes).

As a next step, we aggregated the survey data from Phase 1 and system log data from Phase 2 for each participant. Survey data helped us explore how participants' characteristics, such as familiarity with the foreign language, frequency of using Facebook and MT influenced the ways participants perceived and used the system. In addition, system log data allowed us to investigate how characteristics of posts (e.g., post language, poster type, presence of photo/link) affected participant's use of the system. By incorporating the survey and system log data into consideration, we iteratively validated, compared and synthesized the themes.

4 FINDINGS

In this section, we first present a descriptive overview of how participants used SenseTrans. We then describe how participants used the annotations in their sensemaking processes (**RQ1**). Next, we examine the factors that shaped their perceptions of the usefulness of the different kinds of annotation (**RQ2**). Last, we investigate overall effects of having access to the annotations (**RQ3**).

4.1 Descriptive Overview of Use

On average, participants used SenseTrans for about 13 days, activating it on average for 33 posts (Min=5, Max=116, SD=28.3). Overall, participants used it 565 times on posts in 29 different languages including Spanish, Japanese, Arabic, Hindi, Korean, and Italian. Most of the time (74%, 418 posts), participants used it for foreign language posts that they could not read at all, although sometimes they used it on posts in their native language (8%, 46 posts) or a second language they spoke (18%, 101 posts). About 90% of the posts that participants launched SenseTrans for contained at least one image and half of them included links.

Of posts that participants launched SenseTrans on, participants clicked 'emotion analysis' for 65.1% and 'keyword analysis' for 54.2% of them. When they clicked keyword analysis, 12.4% of time they clicked at least one of the keywords to access the description of those words from Wikipedia. For 25.5% of posts, they just launched the system and used the translation without opening either the emotion or the keyword analysis.

Usage patterns of SenseTrans varied by participant (Table 1). Half of the participants (9/18) checked both the emotion or keyword analysis results most of the time (78.9% for emotion analysis, 81.8% for keyword analysis). Four used emotion analysis much more than keyword analysis (84.5% versus 21.3%), while three used emotion analysis somewhat less than keyword analysis (35.5% versus 50.6%). The other three users rarely used either feature (less than 15% of the time).

Table 1 Participants and System Usage Overview

ID	Sex	Native Language	Second Language	# of posts using SenseTrans	% of those accessing Emotion analysis	% of those accessing Keyword analysis
P1	M	Russian	English, French	49	51%	65%
P2	F	Korean, English	Chinese	21	95%	95%
P3	F	English	Hindi, French	59	100%	23%
P4	F	English	French	24	75%	87%
P5	M	Chinese	English	5	20%	40%
P6	F	English		18	11%	11%
P7	M	English	Spanish, Hebrew	18	38%	50%
P8	F	English	Chinese	116	93%	77%
P9	F	Chinese	English, Spanish	18	94%	38%
P10	M	Chinese	English	28	14%	14%
P11	F	Chinese	English, Spanish	17	58%	52%
P12	F	English	Chinese, Korean	75	66%	92%
P13	F	Chinese	English	58	69%	19%
P14	F	English	Spanish	6	83%	50%
P15	M	English		35	71%	97%
P16	M	English	Spanish	16	100%	100%
P17	M	English		15	0%	13%
P18	M	English	French, Bulgarian	21	47%	61%

4.2 How Did People Use the Annotations in Sensemaking?

Most participants reported that their sense making processes were improved by the emotion and keyword annotations. However, the level of dependency on annotations varied. Depending on how much information participants were able to process or glean from other cues in a post, annotations played primary, supplementary, or confirmatory roles in their sensemaking processes.

When participants could not process, and extract much cues from either the post itself or the MT output, they often relied on emotion and keyword annotations as their primary source of information. In particular, while MT was still the most essential source in their sensemaking processes, many participants felt that MT output often could not suffice to understandings of foreign language posts due to its errors and inconsistencies. For example, P7 described the difficulties in making sense of posts when the translation did not make sense:

> *I'd still read the translation first to understand what it is, but [...] [Translation] didn't seem very fluid but seemed like it was literally transplanting each word, but not in the context of the sentence.* (P7, U.S., Male)

In addition to unintelligible MT outputs, when there were not many other clues that they could glean from such as images or emojis, participants often saw annotations as the only available resource for them, as illustrated by P3:

> *Um, in terms of this one, I just felt like I had to believe it because I had no other choice.* (P3, U.S., Female)

When participants had some sense of a post based on the translation, visual cues, or background knowledge about the poster and topics, they tended to use the analysis results as a

supplementary resource that could complement their partial understanding of a post. As intended in the design of SenseTrans, the annotations provided helped them fill in gaps caused by the cultural and contextual challenges in sensemaking of foreign language posts. For example, P2 had thought that it was almost impossible to gain emotional understanding from a foreign language post without knowledge about the language and its context. However, the emotion analysis allowed her to understand the emotional components of a post to some degree:

> *For emotion analysis, you can't really get that online, unless you understand the language and understand the context of words, but I don't. So, emotional analysis is really handy because the emotion is hard to assess* (P2, Korea, Female)

Likewise, P17 valued the keyword analysis because it could back him up when he had no background knowledge about other countries and cultures:

> *It's about the cultural difference. If it is someone who's prominent in Italy, because I don't live in Italy I don't know who they are. Then, the keyword [analysis] is helpful.* (P17, U.S., Male)

When there were clear visual cues, higher fluency in the language, or excellent quality translation in a post, participants believed that they had mostly made sense of the post by themselves. Still, they sometimes used the annotations to confirm their understanding, comparing the annotations to their own understanding to give them more confidence in and evidence for their sensemaking:

> *Even if the translation makes sense to me, I take a look at the emotion analysis to see whether it is consistent with my judgments.* (P5, China, Male)

While SenseTrans was designed primarily to improve accuracy and depth in sensemaking of individual posts, some participants utilized annotations mostly for skimming or browsing their feed quickly to get the gist of posts and assess their potential interestingness. For example, P6 stated that she usually checked the analysis results first because it was not possible for her to extract any cues for estimating the potential interest of the posts just by glancing at them:

> *If it's something I know [the language of the posts], I can piece some words a bit. I'll have enough idea of whether I want to read it or not. Whereas if it's totally unknown language, there's no way for me to glance at it and know what it's about. So, I'm more inclined to click the analysis for those posts.* (P6, U.S., Female)

Interestingly, some users felt that being able to check the overall slant or key words quickly could be useful even in one's own native language, particularly for longer messages where they might want to figure out the gist of the contents quickly:

> *It does help when the [English] text is super long and something about politics or ideologies. Then, it was interesting to see what sentiment and keywords they had because it gave hints if I would be really interested in what the contents was about.* (P12, U.S., Female)

4.3 What Factors Shaped People's Perceptions of the Usefulness of the Annotations?

Turning to our second research question about what factors shaped participants' perception of the annotations, the interviews revealed three main themes. First, participants iteratively assessed the credibility of annotations by comparing them to their own interpretations. When they matched, this served to increase their confidence not just in their own interpretations, but also in the annotations themselves. Second, they spent time trying to understand how the annotations were generated in order to decide both what they meant and how much to trust them. Finally, the usefulness of different kinds of annotation was affected by participants' goals when browsing their social media newsfeed.

4.3.1 Assessing the Credibility of the Annotations

As participants started using the system, they tended to assess the quality of the annotations by making comparisons between their own assessments and system outputs. For instance, P3 described using SenseTrans on posts written in her second language (Hindi), not because she needed further information, but because she just wanted to assess whether the algorithms came to the same interpretations as she did. After she confirmed that the analysis results were fairly consistent with her judgements, she concluded that the annotations were credible.

> *I used it few times for Hindi posts. At first, I just doubt it because I am not sure. But, it has been consistently correct, so I thought there's some credibility* (P3, U.S., Female)

However, when participants' interpretation did not match the annotations, they judged the credibility of annotations negatively. For example, P7 saw a post with an image where his Vietnamese friend was singing with other people. From the image, he assumed that this post was positive, but the emotion analysis results said this post was 'moderately negative' and contained 'sadness' emotion, which raised doubts:

> *I don't know if it's negative and sad. They look fine and she's singing [in the picture]. I'm not too sure about emotion. I wouldn't know when to trust it. […] I would trust my instinct over the emotion analysis.* (P7, U.S., Male)

Sometimes participants worked to understand and resolve the discrepancies. For example, P11 noticed that the emotion analysis outputs for posts from the Huffposts news source were almost always "neutral" regardless of topic. She originally assumed that sentiment analysis would reflect the valence of the events the post addressed (e.g., an article about threats should carry negative sentiment), and so at first thought that the "neutral" interpretation was not correct. However, after she considered journalistic norms of neutrality, she revised her opinion:

> *I would say it's a little bit off, because this is about threats. But, because it's news, I guess, it's natural to be neutral, because it's for reporting the story not about whether it's happy or not.* (P11, China, Female)

4.3.2 Building Mental Models of the Underlying Algorithms

In addition to assessing credibility of annotations, some participants tried to understand how the annotations worked. This was less for the purpose of gaining a detailed technical understanding of the algorithms; rather, participants wanted to build mental models about how and what the annotations were trying to represent. When they could build a model consistent with the algorithm's behavior, they felt they understood it better and that it was more credible and in turn useful.

For example, many users felt difficulties having a clear understanding of what the keyword analysis considered 'key words'. Operationally, SenseTrans highlighted named entities that were well-known enough to be listed in databases such as Wikipedia or DBpedia that could be used to provide additional information [23]. However, this was not always obvious to participants. For instance, P11 initially expected that the keyword analysis would catch every proper noun, but after using the tool on a number of posts she came to realize that it only highlighted relatively notable public figures, locations, or organizations. Then, based on this understanding, she could decide that the keyword analysis would be more useful for more informational types of posts rather than personal status updates:

> *For the posts from the public pages and celebrities accounts, there would some information that the keyword analysis becomes useful. But in personal accounts, people are just posting their daily lives, so that wouldn't involve a lot of key words.* (P11, China, Female)

However, it was not easy to establish the appropriate mental model of what kinds of words the keyword analysis would highlight even after a few iterations of reasoning about it. For example, P3 first assumed that it would capture only geographic locations such as Puerto Rico. Then she found that big events or celebrities' names were detected as key words. Then she noticed that the name of her cousin was not detected as a keyword, which did not match with her understanding:

> *Do keywords only mean the big organizations or national geographic? Then, why is 'Hurricane Maria' detected? Because it has been big in the news? That's why I don't understand how the keywords work. […] I was wondering why it didn't extract everything like the name of the person.* (P3, U.S., Female)

When people did form wrong mental models, it led to problems. For instance, SenseTrans computes the sentiment analysis based on the text of the original post, but P16 believed the sentiment analysis aggregated others' reactions (likes and comments) toward the post. Because his mental model was far from how it really worked, he felt still confused about how it worked and what it was for:

> *I click on the emotion analysis to kind of see whether it was more towards the right or the left. That one's pretty neutral, and this article just came out 10 minutes ago, so I guess maybe overtime, if more people read it and comment on it, then maybe the emotion would shift. […] I didn't really fully understand how the emotional analysis part worked. What exactly was this for?* (P16, U.S., Male)

4.3.3 Information Needs Based on Social Media Use Goals

As noted in the descriptive overview, some participants were more likely to use one kind of annotation than the other. Issues with credibility and mental models about the different annotations partly drove these preferences, but another factor was whether the kinds of information offered matched people's social media browsing goals. For example, P4 thought that keyword analysis was useful because she used Facebook mostly to keep up with news articles, and saw the emotion analysis as less useful even though she thought it was interesting:

> *I typically see what's in the news. That's how I mostly use Facebook. I thought the part where it gives the Wikipedia was really helpful. Because sometimes I look at something and go look it up myself, but it showed you exactly what it was. […] The sentiment was cool, but, for news articles, it didn't really make much of a difference.* (P4, U.S., Female)

In contrast, P3 was more interested in catching up with her friends and family members on Facebook. She thought the emotion annotations were useful because they helped her grasp the sentiment and emotion of friends' posts more quickly and accurately. She found the keyword analysis less clearly useful for these goals, and thus less useful overall:

> *[My main purpose to use SenseTrans was] to understand what my friends are saying when they write it in different types of languages. It just gives you a basic idea whether somebody's happy with the post, or sad. I like that. […] The keyword analysis, I didn't really understand the purpose of that, so I would just get rid of it.* (P3, U.S., Female)

Finally, some participants had little need for or interest in making sense of foreign language posts, and thus were not very active in using SenseTrans. This happened when they had little foreign language content in their feed, or when their main goal was for entertainment purposes that did not require either deep sensemaking or active filtering. such as browsing for food pictures as described by P10:

> *What attracts me most is the food [related posts]. I just want to get some feeling great, and enjoyable by seeing the food […] I almost ignore the sentiment and key words analysis. Most*

of the posts were just very short text about what is this food. I just give a glance and I will understand the meanings. We are not, like, stupid machine, right? (P10, China, Male)

4.4 What Were the Overall Outcomes of Using the Annotations in a Real Setting?

Although some participants found SenseTrans not to be useful, on balance they reported several useful outcomes, including an increased sense of self-efficacy around sensemaking in foreign language posts, an increase in their awareness of other cultures, and improvements in their relationships with connections who speak other languages.

The most consistent effect of using SenseTrans was that participants reported increased self-efficacy in comprehending foreign language posts. They appreciated that they could make sense of foreign language posts without much additional cognitive effort or specific knowledge of the other language. As a result, they felt they were much more likely to engage with foreign language posts:

If you are not familiar with the language, you are not going to look at the post because you cannot understand it anyway, so you get discouraged. But, now I'm more motivated to look at other posts that are not English. Because I know that I can understand what they are saying at a greater ease. It's just as I'm reading another English post. (P3, U.S., Female)

Some participants reported that sensemaking with SenseTrans could help them gain increased cultural understanding, as with P7's understanding the underlying reason for his friend's Halloween costumes after reading the blurb about "Che Guevara" provided by the keyword analysis:

If you are not from the specific culture, it's difficult to understand. Like with my friend's post, I really don't know who the person is so I don't know the context of what they are talking about or why they are dressed that way. [...] It helps me understand better why. She's Colombian from Cuban, so it makes sense, she would dress up as Cuban iconic that's part of her heritage. (P7, U.S., Male)

Some participants also reported that SenseTrans could further their relationships with connections who often posted in foreign languages, by helping them get to know their foreign connections better:

Without having this installed, it just seemed like she [my Korean friend] was posting really random content that I had absolutely no idea about. [...] Now I know what kind of content she's been posting. (P12, U.S., Female)

This included people whose families extended across countries and languages, who might not otherwise know much about remote family members. During the deployment, P15 revisited earlier posts from his extended family members in Greece that he had not attempted to pay attention to before because of the language difference. The annotations helped him understand their posts better, leading to stronger feelings of connectedness with his extended family members overseas:

Before I used this, if I saw something in a different language, I wouldn't even attempt to translate it, because it was gibberish for me. [...] It was good to go back and find out what family members were saying. I felt I get to stay more connected to my family [in Greece] (P15, U.S., Male)

5 DISCUSSION

In this section, we return to each of our three research questions in more detail. We first summarize the key findings around that question, then discuss what those findings mean for other researchers, ranging from the specific area of cross-language sensemaking in social media to the more general problem of designing intelligent support tools for sensemaking.

5.1 A Better Understanding of Tool Support for Cross-Lingual Sensemaking in Social Media

First, our findings suggest that tools to annotate foreign language posts such as SenseTrans are used in several different ways (**RQ1**). We found that annotations played primary, supplementary, or confirmatory roles in one's sensemaking depending on how much information participants could extract from other aspects of the post. When people had relatively few cues, annotations acted as primary or supplementary sources to bridge gaps in understanding. Even when they had enough cues to understand a post, they used the annotations to help confirm them. On balance, participants valued that SenseTrans allowed them to gain emotional and contextual understandings about a post in-situ with minimal effort.

Because annotations played different roles depending on a person's sensemaking needs with respect to a particular post, suggests there may be value in adaptive designs that provide different types or amounts of annotation based on context. It would be interesting to develop models to estimate how well a given reader could make sense of a given post, using factors our participants identified such as language fluency, the presence of visual or attentional cues, the estimated quality of the translation (for which algorithms exist, e.g., [6][8]), post length, the relationship with the poster, and so on. Predictions from the models could be used to provide different amounts and kinds of sensemaking cues for a given post, with the goal of maximizing benefit while minimizing cognitive load.

In addition, we found that annotations helped some participants quickly get the gist of posts for overall tone and topic as they skimmed their feeds, helping them determine the potential interestingness of each post – for some participants, even in languages they understood. This suggests that in browsing, skimming and filtering the newsfeeds with huge amounts of status updates, the annotations could be the salient and effective cues people could use to complement markers of others' attention such as Likes and Comments.

The finding about getting the gist of messages rather than a deep interpretation has interesting design implication for cross-lingual sensemaking in social media that operates at the level of feeds rather than the level of individual posts. This might suggest potential design changes in the systems like SenseTrans; for instance, annotations need to be visible by default (not 'on-demand' fashion) and provide more intuitive and simpler types of information that users can process quickly. It might also involve deeper integration of these features into the platform rather than bolting on a browser extension. Facebook, for instance, has the resources needed to extract and cache extensive annotation cues and to smoothly embed the results into the post itself, reducing both the computational and cognitive costs of providing annotations. They could also investigate designs aimed at helping people get see relationships between sets of posts, providing in-contexts ways to filter or highlight posts with certain kinds of annotations (e.g., emphasizing positive posts or grouping posts by language).

5.2 Understanding How People Made Sense of the Sensemaking Tool Itself

Our findings also identified three main factors that shaped how people formed credibility and utility judgments of the unfamiliar annotations (**RQ2**). First, participants assessed the credibility of annotations by comparing them to interpretations they were confident in; when these matched, perceived credibility increased, and vice versa. Second, they tried to build mental models to figure

out what annotations represented and how they were generated. When they had unclear or wrong mental models, they usually concluded that the annotations were less useful and credible. Third, they tended to prefer annotations that provided cues that supported their information needs and Facebook use goals. Participants with high information seeking motives tended to prefer keyword to emotion analysis, while those with high relationship maintenance motives preferred emotion to keyword analysis; people who had little need for foreign language sensemaking found neither particularly useful.

On the issue of credibility evaluation, designers need to account for the fact that classification algorithms inevitably make some low-confidence predictions and errors. Many such algorithms, including the sentiment analysis API used by SenseTrans, provide confidence estimates that could be leveraged in design. Perhaps SenseTrans, and other similar systems, might be more effective if they choose not to provide low-confidence annotations, especially early on when people are forming impressions of the system [36]. When it is useful to present lower-confidence predictions, they might include confidence information as a kind of meta-annotation to help people better assess individual annotations [27] and to understand more broadly that the annotations are imperfect and 'seamful' [10]. These tactics could help users make more nuanced trust judgments—though this raises risks that too much information could overwhelm users and decrease trust [21].

SenseTrans also could have done more to help people form appropriate mental models. We provided brief descriptions of the keyword and emotion analyses to users when we set up the extension on each participant's laptops, but these were not sufficient for at least some users to understand how they were computed or what they meant.

This issue is often framed in terms of "explaining" AI algorithms, which has a long history in recommender systems [16][37] and is currently a hot area of research around machine learning more broadly [13]. Our findings suggest considerations for designers around helping people form mental models. Just as people will have different sensemaking goals, abilities, and needs around the *domain* of the system (e.g., cross-lingual sensemaking), the same will be true *with respect to* the system itself. People will have widely varying expertise with the underlying technologies, and willingness to learn about them; this implies favoring designs where people can progressively deepen their inquiry into the system's workings. Further, most existing work on explaining intelligent systems focuses on explaining individual recommendations or decisions. However, our results suggest that people had trouble forming their models by looking at individual instances, suggesting the need to design explanations at a level above individual results.

5.3 Supporting Better Sensemaking, Learning, and Relational Outcomes

Finally, our findings suggested there could be valuable outcomes from using annotations to support cross-lingual sensemaking in social media (**RQ3**). Most participants reported that they experienced positive effects such as increased self-efficacy around sensemaking of foreign language posts, which encouraged them to engage more with such posts. This could in turn help to gain some awareness of other cultures and feel closer to connections who speak other languages.

Taken together, our findings suggest that tools to annotate foreign language posts such as SenseTrans could benefit people's sensemaking processes and outcomes. Previous studies characterized the three main hindrances for sensemaking of foreign language posts as extra cognitive load, lack of knowledge, and lack of relevance [24]. Our results showed that annotations could help users to overcome such hindrances to some extent. Providing extra information in situ helped address the issue of extra cognitive load and lack of knowledge factors. Further, it sometimes addressed the issue of lack of relevance, the perception that posts in other languages were not supposed to be understood by them, by increasing their attention and engagement with foreign language posts.

One concern is that, although we see increased feelings of self-efficacy as overall positive, there is the potential for over-reliance on or over-confidence in the annotations, especially (a) after the initial credibility-testing phase and (b) in situations where they play a primary role because of a relative lack of other information to support sensemaking. Earlier suggestions for estimating sensemaking ability and providing confidence information to encourage credibility assessment may help address these concerns, but designers should broadly consider the risks of encouraging dependence on annotations.

We also see the potential for further improving people's outcomes by diversifying the kinds of sensemaking support offered. For example, SenseTrans could provide the information about pragmatic dimensions of the posts such as the degree of politeness, indirectness, and the existence of humor and sarcasm, which have strong cultural elements, making them interesting problems for cross-cultural sensemaking. Systems could also use computer vision techniques such as object recognition and automatic generation of image descriptions to annotate posts with information extracted from visual components (e.g., [26][40]). As with the pragmatics problems above, the recognition of culturally meaningful objects in images and sentiment analysis of images are interesting research questions in their own right; progress there could provide additional support for the cultural and relational potential offered by cross-language sensemaking in social media.

5.4 Limitations

It is also important to note limitations of the study that might point to interesting future research directions. First, our participants came from a U.S. undergraduate population who spoke English predominantly, leaving open questions of how wider populations would come to perceive and use the system. Second, although the design rationale for SenseTrans was grounded in real cross-lingual sensemaking needs, it is only one point in a much wider design space of annotations and interfaces that is ripe for exploration. Third, as systems with intelligent components are increasingly deployed, people will start to build expectations of how they work and what they offer; this might in turn affect how they form credibility and usefulness judgments around future systems that use similar techniques. Finally, our findings are based on a relatively short-term deployment and our instruction in the study design, might have altered the ways people usually consumed and made sense of social media posts. Therefore, it would be useful to conduct longer-term and larger-scale studies with a range of methods for assessing outcomes.

6 CONCLUSION

The goal of this paper is to investigate how computational tools to annotate communication can support multilingual sensemaking on social media via an in the wild field evaluation study of SenseTrans. Our findings indicated that providing additional annotation could help people's sensemaking processes and outcomes by helping them increase their understanding of and quickly get the gist of foreign language posts. This, in turn, made them more confident about and willing to engage with those posts, providing them with opportunities to gain deeper cultural learning and increased connectedness with their foreign connections. However, these benefits are only likely to be realized when participants formed appropriate credibility judgments and effective mental models of the annotations, and when the annotations matched their social media use goals. Overall, our findings demonstrate the value of designing to support cross-lingual communication and inform design implications for the tools that support cross-lingual communication and sensemaking with algorithmically generated information.

ACKNOWLEDGMENTS

This work was partially supported by National Science Foundation, under grant IIS-1421929 and grant IIS-1910147. The paper was written while Dan Cosley was serving at NSF; it presents the

views of the authors and does not necessarily reflect the views of NSF. We thank Masha Rosca and Vicky Chou for helping conduct the study and Danielle Zeller and Sophie He for helping analyze data.

REFERENCES

[1] Jannis Androutsopoulos. 2015. Networked multilingualism: Some language practices on Facebook and their implications. *International Journal of Bilingualism* 19, 2: 185–205.
[2] Amir Asgarian. 2013. The Iranian EFL teachers' perceptions on translation strategy use. *Procedia-Social and Behavioral Sciences* 70: 922–928
[3] Will Baker. 2012. From cultural awareness to intercultural awareness: Culture in ELT. *ELT journal* 66, 1: 62–7
[4] Kalina Bontcheva and Dominic Rout. 2014. Making sense of social media streams through semantics: a survey. *Semantic Web* 5, 5: 373–403.
[5] Kerstin Borau, Carsten Ullrich, Jinjin Feng, and Ruimin Shen. 2009. Microblogging for language learning: Using twitter to train communicative and cultural competence. In *International conference on web-based learning*, 78–87.
[6] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.
[7] Herbert H Clark. 1996. *Using language*. Cambridge university press.
[8] George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, 138–145.
[9] Irene Eleta and Jennifer Golbeck. 2012. Bridging languages in social networks: How multilingual users of Twitter connect language communities? *Proceedings of the American Society for Information Science and Technology* 49, 1: 1–4.
[10] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 153–162.
[11] Ge Gao, Hao-Chuan Wang, Dan Cosley, and Susan R Fussell. 2013. Same translation but different experience: the effects of highlighting on machine-translated conversations. In *Proceedings of the sigchi conference on human factors in computing systems*, 449–458.
[12] Ge Gao, Bin Xu, David C Hau, Zheng Yao, Dan Cosley, and Susan R Fussell. 2015. Two is better than one: improving multilingual collaboration by giving two machine translation outputs. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 852–863
[13] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*.
[14] Scott A. Hale and Irene Eleta. 2017. Foreign-language Reviews: Help or Hindrance?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17). ACM, New York, NY, USA, 4430-4442.
[15] Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)* 4, 1: 5.
[16] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 241–250.
[17] John Hutchins. 2009. Multiple Uses of Machine Translation and Computerised Translation Tools. *Machine Translation*: 13–20.
[18] Rieko Inaba, Yohei Murakami, Akiyo Nadamoto, and Toru Ishida. 2007. Multilingual communication support using the language grid. In *Intercultural Collaboration*. Springer, 118–132.
[19] Eun-Young Kim, Sun-Mih Park, and Sun-Hye Baek. 2011. Twitter and implications for its use in EFL learning. *Multimedia Assisted Language Learning* 14, 2: 113–137.
[20] Nikiforos Karamanis, Saturnino Luz, and Gavin Doherty. 2011. Translation practice in the workplace: contextual analysis and implications for machine translation. *Machine Translation* 25, 1: 35.
[21] René F Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395.
[22] Claire Kramsch. 2006. The multilingual subject. *International Journal of Applied Linguistics* 16, 1: 97–110.
[23] Hajin Lim, Dan Cosley, and Susan R Fussell. 2018. Beyond Translation: Design and Evaluation of an Emotional and Contextual Knowledge Interface for Foreign Language Social Media Posts. In *Proceedings of the 36th Annual ACM Conference on Human Factors in Computing Systems* (CHI '18).
[24] Hajin Lim and Susan R Fussell. 2017. Making Sense of Foreign Language Posts in Social Media. *PACM on Human-Computer Interaction* 1, 2.
[25] Hajin Lim and Susan R. Fussell. 2017. Understanding How People Attend to and Engage with Foreign Language Posts in Multilingual Newsfeeds. In *Eleventh International AAAI Conference on Web and Social Media*.
[26] Julian McAuley and Jure Leskovec. 2012. Image labeling on a network: using social-network metadata for image classification. In *European conference on computer vision*, 828–841

[27] Sean M McNee, Shyong K Lam, Catherine Guetzlaff, Joseph A Konstan, and John Riedl. 2003. Confidence displays and training in recommender systems. In *Proc. INTERACT*, 176–183
[28] Alireza Moghaddam. 2006. Coding issues in grounded theory. *Issues in educational research* 16, 1: 52-66.
[29] Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. 2014. Information network or social network?: the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, 493–498
[30] Joyce S. Osland and Allan Bird. 2000. Beyond sophisticated stereotyping: Cultural sensemaking in context. *The Academy of Management Executive* 14, 1: 65–77.
[31] Emrah Özdemir. 2017. Promoting EFL learners' intercultural communication effectiveness: a focus on Facebook. *Computer Assisted Language Learning* 30, 6: 510–528
[32] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, 2–4
[33] Louise J Rasmussen, Winston R Sieck, and Joyce Osland. 2010. Using cultural models of decision making to develop and assess cultural sensemaking competence. *Advances in Cross-Cultural Decision Making. Boca Raton, FL: CRC Press, Taylor & Francis Group*
[34] Tomohiro Shigenobu. 2007. Evaluation and usability of back translation for intercultural communication. *Usability and Internationalization. Global and Local User Interfaces*: 259–265
[35] Tomohiro Shigenobu, Kunikazu Fujii, and Takashi Yoshino. 2007. The role of annotation in intercultural communication. In *International Conference on Usability and Internationalization*, 186–195
[36] Kirsten Swearingen and Rashmi Sinha. 2002. Interaction design for recommender systems. In *Designing Interactive Systems*, 312–334.
[37] Nava Tintarev. 2007. Explanations of recommendations. In *Proceedings of the 2007 ACM conference on Recommender systems*, 203–206.
[38] Hao-Chuan Wang, Dan Cosley, and Susan R Fussell. 2010. Idea expander: supporting group brainstorming with conversationally triggered visual thinking stimuli. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 103–106.
[39] Charles Wankel. 2016. Developing cross-cultural managerial skills through social media. *Journal of Organizational Change Management* 29, 1: 116–124
[40] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1180–1192
[41] Naomi Yamashita, Rieko Inaba, Hideaki Kuzuoka, and Toru Ishida. 2009. Difficulties in establishing common ground in multiparty groups using machine translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 679–688
[42] Naomi Yamashita and Toru Ishida. 2006. Effects of machine translation on collaborative work. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, 515–524
[43] Takaya Yuizono, Wei Li, and Jun Munemori. 2011. Promoting cultural learning: Effects of cultural knowledge on text chats between Japanese and Chinese participants. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, 167–176
[44] Dejin Zhao and Mary Beth Rosson. 2009. How and why people Twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, 243–252
[45] Jichang Zhao, Junjie Wu, and Ke Xu. 2010. Weak ties: Subtle role of information diffusion in online social networks. *Physical Review E* 82, 1: 16105
[46] See Translations for Posts and Profiles | Instagram Help Center. Retrieved April 02, 2019 from https://help.instagram.com/249420315434256
[47] ATLAS.ti: The Qualitative Data Analysis and Research Software. Retrieved from http://atlasti.com/

Received April 2019; revised June 2019; accepted July 2019