Annotation Process for the Dialog Act Classification of a Taglish E-commerce Q&A Corpus

Jared Rivera, Jan Caleb Oliver Pensica,
Jolene Valenzuela, Alfonso Secuya, Charibeth Cheng
De La Salle University, Manila, Philippines
{jared_rivera, jan_pensica, jolene_valenzuela,
alfonso_secuya, charibeth.cheng}@dlsu.edu.ph

Abstract

With conversational agents or chatbots making up in quantity of replies rather than quality, the need to identify user intent has become a main concern to improve these agents. Dialog act (DA) classification tackles this concern, and while existing studies have already addressed DA classification in general contexts, no training corpora in the context of e-commerce is available to the public. This research addressed the said insufficiency by building a text-based corpus of 7,265 posts from the question and answer section of products on Lazada Philippines. The SWBD-DAMSL tagset for DA classification was modified to 28 tags fitting the categories applicable to e-commerce conversations. The posts were annotated manually by three (3) human annotators and preprocessing techniques decreased the vocabulary size from 6,340 to 1,134. After analysis, the corpus was composed dominantly of single-label posts, with 34% of the corpus having multiple intent tags. The annotated corpus allowed insights toward the structure of posts created with single to multiple intents.

1 Introduction

An essential part of social media is the messaging feature which is easily adopted due to its convenience and speed in comparison to other communication methods (Alison Bryant et al., 2006). In the Philippines, most online sellers prefer using social media as an e-commerce platform for their businesses for exactly this reason (Marcelo, 2018). However, for social media to be effective as an e-commerce platform, active participation of the seller and the customer in the conversation is required. A general drawback in e-commerce is the lack or unavailability of sales clerks (i.e. online shop moderators) to interact with customers online. This problem is commonly evident among solo retailers, which provides an opportunity for the use of conversational agents to act as sales clerks on behalf of these sellers (Bogdanovych et al., 2005). Taglish (Tagalog-English) is often comfortably used on Philippine social media, and since natural language is noted as the most natural means of communication between humans (Weischedel et al., 1989), interaction using Taglish as natural language, is seen as a feasible option to connect these conversational agents with Filipino customers online (Hill et al., 2015).

However, misunderstandings are common in conversational interactions, more so when an online platform is used and transactions are conducted online than in-person, and the problem may be more complex once a machine is on one end of the conversation. Despite its capability to participate in a conversation, conversational agents still fail to simulate and capture the essence of the full range of an intelligent human conversation (Hill et al., 2015). The identification of dialog acts in an utterance is therefore an important goal of any system aiming to properly establish intent among participants to understand a conversation.

In this work, we focus on collecting and annotating the dialog acts of queries within the domain of e-commerce, specifically from Lazada Philippines, building a corpora with dialog act annotations named LazadaQA-Taglish-7k. The dataset is open sourced in a public repository. [1]

2 Related Works

A dialog act (DA) represents the intention of a person's utterance (Austin and Urmson, 1962). According to Stolcke et al. (2000), DAs may be considered as a set of tags that classifies utterances according to a combination of pragmatic, semantic, and syntactic criteria. In addition, it is de-

[1] https://github.com/dlsucelt/lazadaQA

Proceedings of the Second Workshop on Economics and Natural Language Processing, pages 61–68
Hong Kong, November 4.

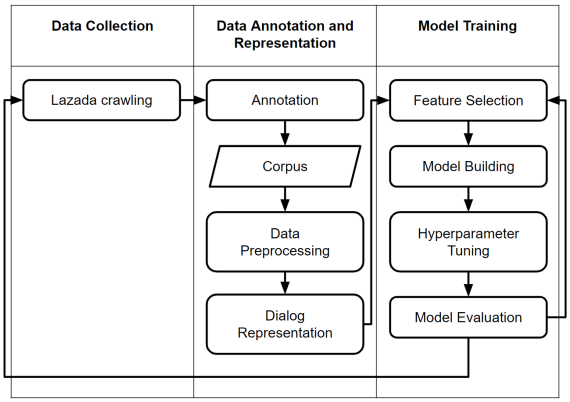

Figure 1: Flowchart of the Study

scribed to be a useful first level of dialog understanding to describe the structure of a conversation. There are four (4) commonly used publicly-available corpora that are usually used for training in DA classification: Switchboard (Godfrey et al., 1992), MapTask (Anderson et al., 1991), MRDA (Janin et al., 2003), and VERBMOBIL (Wahlster, 1993). It is noticeable that among all four corpora mentioned, there are no works that are applicable to the e-commerce setting. As of the time of writing there is only one e-commerce related work on DA classification by Meng and Huang (2017), which used a proprietary Chinese conversational dataset from a Chinese e-commerce service, however the dataset is not publicly available and details regarding its data collection were not specified. The lack of data for e-commerce dialogs motivated the building of the corpus for this work.

3 Methodology

The structure of the methodology for this study is illustrated in Figure 1. It is mainly divided into three phases, namely Data Collection, Data Annotation and Representation, and Model Training. Only the first two phases will be discussed, while the third phase will be briefly tackled in Section 5. Subsections that describe the steps per phase in detail follow.

3.1 Data Collection

For this work, 7,265 posts were scraped from the Q&A sections of products under the categories electronic devices and appliances, namely mobile phones, laptops, printers, and peripherals. These categories were chosen because of the nature of electronics which consists of many variation of components and specifications that can possibly lead to a higher number of inquiries.

The data collection was done in two iterations: The first iteration crawled 1,967 posts under Audio Devices, and Computers and Laptops using Octoparse (Oct, 2018). The following iteration crawled an additional 5,298 posts under Printers, Mobile Accessories, and Audio Devices using a Python script that utilizes Selenium automated testing. Each post contains an utterance from a customer ("question") and a seller ("answer"), customer, seller, and time posted.

The final dataset contains posts from 39 unique sellers with 3,437 instances from the Audio Devices subcategory, 1,021 from Computers and Laptops, 1,365 from Mobile Accessories, and 1,442 from Printers.

3.2 Data Annotation

The annotation of data was done by three (3) individuals in parallel, guided by a list of tags. Final tags assigned to a post were decided by majority agreement such that if 2 up to 3 out of 3 annotators agreed that a post be assigned to a certain category, it will be assigned as such.

Each post could be classified with more than one (1) tag, thus presenting a Multilabel Classification Problem. This was addressed by transforming the labels by Binary Relevance. Labels were added to each post in their actual form (i.e. `Availability Inquiry`) and then converted to a binary vector with the length corresponding to all tags (28), with values 0 or 1 corresponding to whether a tag is applicable to the post.

The tagset used for identifying DAs was initially based on the SWBD-DAMSL tagset by Jurafsky et al. (1997) and then modified based upon common intents found in the posts. This led to the emergence of tags for the context of e-commerce. Tags used for the study are listed on Table 1 accompanied by examples where the tags apply.

The annotation was done in 4 iterations, with the tagset evolving over the course of the iterations. For the first iteration, 1,967 posts were crawled from Audio Devices and Computers and Laptops. The initial content included the initial dialog acts used in the Messenger dataset. `Discount Inquiry` was removed due to its similarity to `Promo Inquiry` in terms of definition.

In the second iteration, the non-occurring

Table 1: Tags used to annotate LazadaQA-Taglish-7k. Translations for Taglish phrases are provided in parentheses.

	Tag	Example
Inquiry	**Availability inquiry**	*Is the iPhone C still available for purchase?*
	Price inquiry	*kano ba IPHONEX?* *("how much is IPHONEX?")*
	Specification inquiry	*so wait.. ano ba features ng samsung* *("so wait.. what are the features of samsung")*
	Contact details inquiry	*Can I have your contact information?*
	Promo inquiry	*and are there any applicable promos that* *can be used for buying phones?*
	Delivery inquiry	*Do you ship?*
	Payment method inquiry	*Hi! I am inquiring about the Razer* *Blade Stealth, what are the means of payment?*
	Definition inquiry	*Ahm, itatanong ko lang haha. Ano yung ibig sabihin* *ng "unlocked"?* *("Ahm, I just wanted to ask haha. What does "unlocked" mean?")*
	Process inquiry	*Hi, nakita ko sa page niyo na may swap or sale* *for electronics, may I ask how the swap system works?* *("Hi, I saw that on your page there is swap or sale* *for electronics, may I ask how the swap system works?")*
	Product recommendation request	*Hi! What phone models do you recommend* *for a mid-ranged budget?*
	Request (action-directive)	*Please meet her near the university.*
	Clarification	*Under mobile networks, right?*
	Warranty inquiry	*wala po talaga sya warranty?* *("it really has no warranty?")*
	Inquiry (others)	*Is this legal, though?*
Complaint	**Service complaint**	*I even contacted you guys so many times already,* *but you guys never answer me properly.* *It is already so frustrating.*
	Product complaint	*Parang may problem ata sa hardware,* *di gumagana yung LTE ng SIM* *("It seems like there might be a problem with the hardware,* *the LTE of the SIM doesn't work")*
	Price complaint	*parang awa niyo na ito ba talaga price nito* *baka naman hindi bat ganun ang total pag* *add ko sa chart ko 1796* *("please is this really the price* *maybe its not why is the total like that* *after I add the price to my chart 1796")*
	Delivery complaint	*Excuse me i ordered iphone X* *bakit bato at sibuyas ang laman!!!!* *("Excuse me i ordered iphone X* *why is it full of rocks and onions!!!!")*
Expression	**Agreement / Accept / Yes-answer**	*Ok that would be fine.*
	Opening	*Hello*
	Thanking	*Ok thank you siz*
	Expression	*Huhu*
Transaction	**Purchase**	*I would like to order one iPhoneX* *through COD please*
	Order cancellation	*pwede ba cancel nlang iba nlang ooderin ko.* *("is it possible to cancel instead I will order something else.")*
	Return / Exchange / Refund	*if ever na may defect sya maam can i return it?* *("if ever it has a defect maam can i return it?")*
Other	**Backchannel**	*Ok wala na po ba tawad yan* *("Ok is there really no discount for that")*
	Follow-up	*wala pang reply ata sa tanong ko?* *("there might still be no reply to my question?")*
	Other	*Uy may nanalo na raw. :O* *("Uy they said someone won already. :O")*

tags `Swap` and `Negative / Reject / No-answer` were removed. In addition, the `Swap` tag was unique to e-commerce conversations on Messenger and is not a feature of Lazada Philippines. `Closing` was also removed due to the annotators experiencing difficulty in classifying such statements and the nature of QA postings being different from a linear conversational flow.

5,298 data points were annotated in the following iterations from Mobile Accessories, Audio Devices, and Printers. The third iteration saw an abundance of posts relating to product warranty, product return, exchange, or refunding, and order cancellations—all of which did not have corresponding tags in the tagset. The following tags were added to the tagset before the next iteration was started: `Warranty Inquiry`, `Price Inquiry`, `Price Complaint`, `Order Cancellation`, and `Return / Exchange / Refund`.

In addition, the tag `Delivery method inquiry` was renamed to `Delivery inquiry` as it was assumed for previous iterations that delivery-related inquiries only ask about possible methods of delivery (e.g. meet-up, courier, pick-up). There were no tags for certain instances of delivery-related inquiries such as asking for the estimated time of delivery, delivery fee, and about specific couriers in the tagset. Instead of adding new tags for each scenario, the tag `Delivery method inquiry` was made into a general tag that encapsulated all delivery-related inquiries.

Lastly, for the final iteration, the `Question (others)` tag was changed to `Inquiry` to be consistent with all inquiry tags.

The final tag distribution can be found in Table 2.

3.3 Inter-rater reliability

The Fleiss κ value, which extends the Cohen κ statistic to more than 2 annotators, was used to measure the inter-rater reliability of the annotators. For this study, the majority agreement was also measured among annotators as to decide the ground truth for DA labeling for the classification task in Section 5. The paradox of high-agreement (majority percentage) and low-reliability (κ value) was found to exist in this case. The computation was done through the following:

Let N be the number of messages, n be the number of annotators, k be the number of dialog act tags, i be the index of messages, j be the index of annotators, and n_{ij} as the number of annotators who assigned the j-th tag to the i-th message. First solve for p_j

$$p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}, 1 = \sum_{j=1}^{k} p_j \tag{1}$$

Table 2: LazadaQA-Taglish-7k Tag Distribution

Tag	Occurrence
Specification Inquiry	4143
Opening	971
Inquiry (others)	684
Thanking	679
Other	396
Product Complaint	389
Delivery Complaint	386
Delivery Inquiry	362
Availability Inquiry	351
Process Inquiry	347
Price Inquiry	265
Expression	175
Request	168
Service Complaint	131
Payment Method Inquiry	107
Warranty Inquiry	101
Return / Exchange / Refund	82
Price Inquiry	65
Contact Details Inquiry	65
Backchannel	61
Definition Inquiry	60
Follow-up	57
Price Complaint	51
Clarification	51
Product Recommendation Request	41
Purchase	33
Order Cancellation	30
Agreement / Accept / Yes-answer	16

where p_j is the proportion of all assignments to the j-th tag. Then compute for P_i

$$P_i = \frac{1}{n(n-1)}[(\sum_{j=1}^{k} n_{ij}^2) - (n)] \tag{2}$$

where P_i shows how many annotator pairs are in agreement for all possible pairs. Next compute for P

$$P = \frac{1}{Nn(n-1)}(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn) \tag{3}$$

where P is the mean of the P_is. Then compute for P_e

$$P_e = \sum_{j=1}^{k} p_j^2 \tag{4}$$

where P_e is the expected mean proportion of agreement. Lastly, plug the values of P and P_e into the following equation to get the value of κ:

$$\kappa = \frac{P - P_e}{1 - P_e} \tag{5}$$

Table 3: Kappa Scores from Highest to Lowest Reliability

	Label	Kappa
1	Opening	0.9276
2	Payment method inquiry	0.9036
3	Thanking	0.8920
4	Warranty inquiry	0.8687
5	Availability inquiry	0.8491
6	Specification inquiry	0.8359
7	Price inquiry	0.7869
8	Product complaint	0.7628
9	Delivery complaint	0.7530
10	Contact details inquiry	0.7249
11	Delivery inquiry	0.7228
12	Promo inquiry	0.6499
13	Price complaint	0.5991
14	Service complaint	0.5296
15	Product recommendation request	0.5035
16	Process inquiry	0.4561
17	Expression	0.3760
18	Definition inquiry	0.3747
19	Order cancellation	0.3076
20	Other	0.2532
21	Backchannel	0.2392
22	Purchase	0.2321
23	Follow-up	0.2233
24	Request	0.2141
25	Clarification	0.1998
26	Agreement / Accept / Yes-answer	-0.0012
27	Return / Exchange / Refund	-0.0051
28	Inquiry (others)	-0.0379

where $1 - P_e$ defines the degree of agreement attainable above chance while $P - P_e$ stands for the actual degree of agreement achieved above chance. There are 2 possibilities for the value of κ, namely: $\kappa = 1$ means complete agreement while $\kappa = -1$ means complete disagreement.

The kappa scores can be found in Table 3.

3.4 Cleaning and Pre-processing

Before the data was used, empty rows were removed from the dataset. All of the posts were converted to lowercase, and terminal punctuation and emojis (digital icons supported by Unicode) were retained as we believe that they were important to the identification of intent behind a post (e.g. angry emoticons may signify complaints). Strings composed of non-separated terminal punctuation and emojis were split by spaces in order to reduce unigrams composed of the same character (e.g. "!!!" turn into '!','!','!'). Text normalization was also applied to the dataset, standardizing all numbers as the token "<num>", stopwords to "<st>", and rare words (words with 0.01% term frequency) to "<rr>". Many rare words normalized related to product titles and details that only occured in a single forum and had no bearing to the intent of the post (e.g. earphones, airpods). This process significantly reduced the vocabulary size from 6,340 to 1,134.

4 Results and Discussion

The results of this study will mainly focus on the analysis of the crawled and annotated dataset, including figures to identify significant observations among the DAs.

4.1 Data Analysis

From Figure 2, while posts annotated with only one tag are dominant within the dataset, 34% of posts within the dataset are still classified under more than one tag, with a significant number of these posts having tag pairs (two tags).

Many observations can be made from Table 4 as to the possible relation between tags. Most of the tags under the inquiry group show similar words that are used in Taglish conversations implying a question (e.g. *"ba"*, a word in the Filipino vocabulary frequently used to ask for clarification) while also having words that relate to each individual tag's intent (e. g. "free" word is frequent among posts tagged as `Promo Inquiry`).

Under the complaint group however, most words are nouns pertaining to the order or item purchased, as topics of the complaint. The word *"lang"* ("only") also appears in complaint tags which may pertain to a lack of or of ill punctuality. There also appears to be an intersection between words used in complaints as well as inquiries, such as words *"lang"* ("only") and *"bakit"* ("why"). This suggests that complaints are often presented in inquiry form. The only tag with unique common words in contrast to the other complaint tags is `Price Complaint`, with many words relating to money such as *"mahal"* ("expensive"), "price", and "fee".

As for expression tags, most words are common among the tags such as "hi", "hello" and "thank". This could mean that most posts are structured to portrait all these intents, and posts that open a conversation could also close with an agreement or thanking expression.

For the transaction group, there is an appearance of words relating to an order or item, and imply a process (e.g. *"paano"* ("how"), "return", "order", "cancel"). While words such as "order" and "item" appear in the complaint group, the presence of expression tags, specifi-

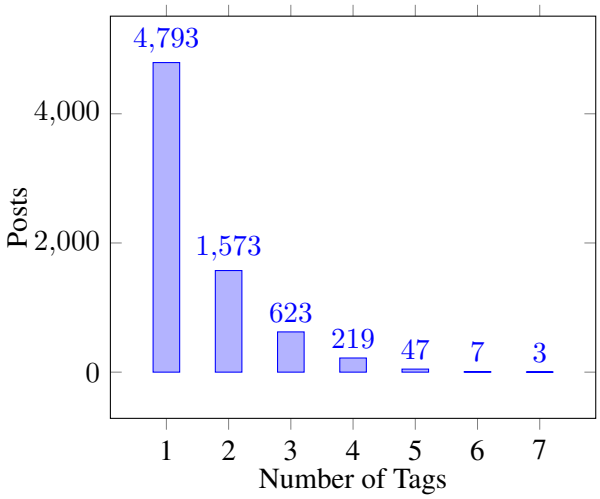

Figure 2: Multilabel Count Distribution

Table 4: Common Words used by Tagged Posts

	Tag	Common Words
Inquiry	**Availability inquiry**	*ba, available, meron, color, stock*
	Price inquiry	*ba, much, shipping, price, magkano, fee*
	Specification inquiry	*ba, pwede, compatible*
	Contact details inquiry	*ba, store, warranty, contact*
	Promo inquiry	*free, ba, shipping, sale*
	Delivery inquiry	*ba, order, day, ilang, delivery*
	Payment method inquiry	*cod, ba, installment, pwede, cash, delivery*
	Definition inquiry	*go, ano, jbl, ba*
	Process inquiry	*ba, order, paano, item*
	Product recommendation request	*pwede, printer, hi, item, one, thank*
	Request (action-directive)	*order, please, ba, item, thank, sana*
	Clarification	*ba, warranty, order, lang, hindi*
	Warranty inquiry	*warranty, ba, item, paano*
	Inquiry (others)	*ba, original, order, lang, bakit*
Complaint	**Service complaint**	*order, item, bakit, naman, wala*
	Product complaint	*lang, item, ba, hindi, bakit*
	Price complaint	*shipping, mahal, fee, price, bakit, lang*
	Delivery complaint	*order, bakit, wala, day, item*
Expression	**Agreement / Accept / Yes-answer**	*thank, ok, yes, opo*
	Opening	*hi, ba, hello, thank, lang, order, pwede*
	Thanking	*thank, order, ba, hello*
	Expression	*order, ba, naman, hindi, thank, sana*
Transaction	**Purchase**	*order, thank, hello, sana*
	Order cancellation	*order, cancel, lang*
	Return / Exchange / Refund	*item, paano, ba, order, return*
Other	**Backchannel**	*ba, naman, sabi, please*
	Follow-up	*hi, order, item, wala, follow*
	Other	*order, item, ba, lang*

cally, `Thanking` may differentiate the negative implications of a complaint from a transaction.

Lastly, tags classified under other, have no identifiable words distinguishable from the other tag categories since these tags hold a broader scope that cannot be properly defined. Tags, `Follow-Up` and `Backchannel`, both require the element of context to properly classify a post as such, while `Other` remains a catch-all tag that is given if a post cannot be classified under any other tag.

5 Application and Current Usage

The resulting corpus and annotations were used to create e-commerce dialog act classifiers. The best-

Table 5: Results from the best DA classification models as an application of this work.

	SVM (BoW)		MLP (TF-IDF)	
	Train	Test	Train	Test
Accuracy	99.46%	99.07%	84.15%	83.58%
Precision	98.03%	96.19%	71.54%	68.26%
Recall	94.17%	89.56%	85.73%	79.71%
F1-score	95.97%	92.68%	75.17%	70.58%

performing machine learning model was a Support Vector Machine (SVM) that used Bag of Words (one-hot encoding) on the questions as features while the best-performing deep learning model was a Multilayer Perceptron (MLP) that used TF-IDF as features. A summary of the results for the best models from this phase of the study can be found on Table 5.

6 Conclusion

This study was able to collect a total of 7,265 posts from the Q&A sections of products in Lazada Philippines. The posts were from products under printers, speakers, and electronic devices and a Python script with Selenium automated testing. The entries contain a question (customer utterance), an answer (seller utterance), the customer, seller, and the timestamp for the post. The corpus was annotated manually by three (3) human annotators using a tagset of 28 dialog acts tailor-fit for e-commerce conversations which were based on the SWBD-DAMSL tagset by Jurafsky et al.. Analysis of the corpus revealed the multilabel nature of posts as well as intersections of common words and intent, within and among the tag groups. Finally, the LazadaQA-Taglish-7k provides a foundation for the use of Taglish in conversational agent interactions as it is the first e-commerce corpora of its kind in its language, which can be applied in the development of conversational agents in the said domain as well as other related fields.

References

2018. Octoparse.

J Alison Bryant, Ashley Sanders-Jackson, and Amber MK Smallwood. 2006. Iming, text messaging, and adolescent social networks. *Journal of Computer-Mediated Communication*, 11(2):577–592.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC Map Task Corpus. *Language and speech*, 34(4):351–366.

John Langshaw Austin and James Opie Urmson. 1962. *How to Do Things with Words. The William James Lectures Delivered at Harvard University in 1955.[Edited by James O. Urmson.].* Clarendon Press.

Anton Bogdanovych, SJ Simoff, Carles Sierra, and Helmut Berger. 2005. Implicit training of virtual shopping assistants in 3d electronic institutions. *e-commerce*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Jennifer Hill, W. Randolph Ford, and Ingrid G. Farreras. 2015. Real conversations with artificial intelligence: A comparison between human—human online conversations and human—chatbot conversations. *Computers in Human Behavior*, 49:245–250.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The ICSI meeting corpus. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–I. IEEE.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard-DAMSL labeling project coder's manual. *Technická Zpráva*, pages 97–02.

Patrizia C Marcelo. 2018. Most filipino merchants prefer social media as top e-commerce platform — survey. *BusinessWorld*.

Lian Meng and Minlie Huang. 2017. Dialogue Intent Classification with Long Short-Term Memory Networks. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 42–50. Springer.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Wolfgang Wahlster. 1993. Verbmobil. In *Grundlagen und anwendungen der künstlichen intelligenz*, pages 393–402. Springer.

Ralph Weischedel, Jaime Carbonell, Barbara Grosz, Wendy Lehnert, Mitchell Marcus, Raymond Perrault, and Robert Wilensky. 1989. White Paper on Natural Language Processing. In *Proceedings of the Workshop on Speech and Natural Language*, HLT

’89, pages 481–493, Stroudsburg, PA, USA. Association for Computational Linguistics.