Building a Swedish Open-Domain Conversational Language Model

Building a Swedish Open-Domain Conversational Language Model

Tobias Norlund
Chalmers University of Technology
Recorded Future
tobiasno@chalmers.se

Agnes Stenbom
Royal Institute of Technology
Schibsted Media Group
astenbom@kth.se

Abstract

We present on-going work of evaluating the, to our knowledge, first large generative language model trained to converse in Swedish, using data from the online discussion forum Flashback. We conduct a human evaluation pilot study that indicates the model is often able to respond to conversations in both a human-like and informative manner, on a diverse set of topics. While data from online forums can be useful to build conversational systems, we reflect on the negative consequences that incautious application might have, and the need for taking active measures to safeguard against them.

1 Introduction

Dialog is an important means through which machines can exhibit intelligence toward humans, which is interesting from a general AI perspective. But dialog also constitutes a natural interface for humans to interact with technology, which opens up for a breadth of applications involving complex information acquisition, automation of tasks and smart support systems. A promising direction towards this goal is the development of open domain conversational systems using large neural networks.

Early approaches to neural conversational systems rely on various forms of Recurrent Neural Networks (RNN) trained autoregressively to model the textual sequences (Shang et al., 2015; Vinyals and Le, 2015; Sordoni et al., 2015; Serban et al., 2016). More recently, as large pretrained Transformer networks have come to dominate progress in NLP in general (Devlin et al., 2019; Radford, 2018; Radford et al., 2019; Brown et al., 2020; Raffel et al., 2020), approaches such as DialoGPT (Zhang et al., 2020), Meena (Adiwardana et al., 2020) and Blender (Roller et al., 2020) have proven the architecture's applicability in open domain dialog systems as well.

However, as the research effort is predominantly put into making progress on English, the importance of making progress in other languages as well has been noted (Ruder, 2020; Wali et al., 2020). Each language is its own unique challenge for many reasons, but the difference in availability of resources is a major one, in particular for data-driven methods. We argue this is also important to keep the public debate on the risks and ethical aspects of large scale language models open to non-English speaking communities. Toward those ends, we present the first (to our knowledge) attempt to build a large scale open domain dialog system in Swedish based on data from Flashback, one of the largest social discussion forums in Sweden. We also present early indicative results on a human evaluation to assess its response generation capabilities across a wide range of topics.

2 Data and preprocessing

Flashback[1] is a Swedish online forum that launched in 1996 and has since grown to become one of the country's most popular social medias (Internetstiftelsen, 2019). In the various sub forums, a breadth of topics are openly discussed including computers and programming, economics, politics, sports and science. To the general public however, the forum is also widely known for housing an anonymous safe haven for controversial subjects such as prostitution, drugs and conspiracy theories (Östman and Aschberg, 2015). Due to its consistent popularity over the last two decades, it arguably today makes up Sweden's biggest single source of general conversational text.

On Flashback, posts are chronologically organized into threads. In a single thread, the discussion is centered around a specific topic typically

[1] http://www.flashback.org

Number of layers	48
Dimensionality	1600
Feed-forward dim	5400
Number of heads	16
Number of parameters	1.4B
Max context length	400
Batch size	512
Optimizer	Adam
Vocabulary size	52,000

Table 1: Model hyperparameters

described by a thread title. Acknowledging the potential for embedding undesired biases, we have initially chosen to use a complete and unfiltered dump of the forum for this study.

The data was tokenized into strings of BPE tokens (Sennrich et al., 2016) using a customly trained vocabulary. Due to Flashback's organization of posts into a single linear feed (unlike the tree structure on e.g. Reddit), it is common that users quote the previous post they respond to, to avoid confusion. As a quote holds important contextual information to a post, we chose to explicitly include this in the way we formatted the threads. More details of how the data was formatted into strings can be found in Appendix A.

3 Model

Following previous works on open-domain dialogue systems (Zhang et al., 2020; Adiwardana et al., 2020), we trained an auto-regressive language model using a slightly modified Transformer (Vaswani et al.) decoder as proposed by Radford et. al. (2019). That is, for an input sequence of tokens $x_1, ..., x_n$, the language model is trained to maximize the likelihood of the joint probability:

$$p(x_1, ..., x_n) = p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}, ..., x_1) \quad (1)$$

We denote our model *Flashback-GPT*, where GPT is an acronym for Generative Pre-trained Transformer as first coined by Radford et. al. (2019). The hyper-parameters chosen are similar to those of the largest variant of GPT-2 (Radford et al., 2019), and are detailed in Table 1.

The model was trained on 16 Nvidia Tesla V100 SXM2 GPUs for 7 days, equivalent to 86,250 gradient updates. The learning rate was increased linearly for the first 5,000 steps up until $5\mathrm{e}{-5}$, after which it was kept constant. We used the `deepspeed` (Rasley et al., 2020) library to optimize memory efficiency across the devices during training.

4 Evaluation

Evaluating natural language generation systems is known to be hard. Even though it is common to conduct automatic evaluations due to their low cost, a human evaluation often serves as an additional validation of the results. However, designing a human evaluation to measure a specific quantity is also not trivial since there is always room for interpretation among the human annotators.

Therefore, we present a pilot study where the main aim is merely to get early indications rather than definite results, and to guide the design of bigger future studies. We design our pilot to measure our quantity of interest: To which extent is the model capable of participating in social discussion forums across a diverse set of topics?

To that end, we seek to measure two quantities: *humanlikeness* and *informativeness*. As language models can often be inconsistent and show lack of commonsense knowledge, humanlikeness is supposed to answer if there is anything in a response that seems off, suggesting it has not been written by a human. However, a response can be humanlike but still uninformative. The notion of ”informativeness” is particularly interesting in our setting as forums can be relatively knowledge centric, and uninformative responses such as *I don't know* add little to the discussion.

4.1 Study design

The study was designed as follows. We select a set of N Flashback threads, held out from training, to be used in the study. For each thread, we only take the first two or three posts to limit the discussion context. We then, for each thread, swap the last post for an alternative generated by the model. Along with the originals, we now have $2N$ threads that we present (in shuffled order) to human annotators. For each thread, we ask two binary questions to measure humanlikeness and informativeness respectively:

1. Is there any indication that the last message was not written by a human?

2. Do you think that the last message adds information to the discussion?

This draws close resemblance to previous evaluations performed on English systems (Zhang et al., 2020; Adiwardana et al., 2020). In Zhang et. al. (2020), humans are asked to rank two alternative responses according to *informativeness*, *humanlikeness* and *relevance*. In Adiwardana et. al. (2020), humans are instead asked the binary questions whether a response ”makes sense” and also whether it is ”specific”, and the average of the two (Sensibleness and Specificity Average - SSA) is found to correlate with humanlikeness. For simplicity, we chose to directly ask for humanlikeness instead of the SSA proxy questions. The complete annotator guideline (Swedish) is included in Appendix B for reference.

For the pilot study, we collected a sample of $N = 120$ Flashback threads, stratified across 12 of the top level forums. We then formed two groups of human annotators with three persons in each group. Each group was presented 60 threads with generated responses, and 60 original, with no overlap. The threads included were randomly chosen, except for a few criteria that we employed to prevent the annotators from exploiting obvious surface patterns when answering question 1.

- As has been noted previously (Roller et al., 2020), beam search decoding strategies have a tendency to generate shorter responses over longer. We decided to only include threads where the last (human written) response is at most 200 characters.

- Since the model supports a maximum sequence length of 400 tokens, we exclude threads where the context is longer than 350 tokens, to leave some room for the generated response.

- Since the model often fails to generate correct quotes of previous responses, we remove any quotes from the last (human written) response, and force the model not to generate quotes as well.

We include a subset of the threads (both with generated and ground truth responses) in Appendix C.

4.2 Decoding

The decoding strategy used to generate responses from neural language models is an important part of the system as a whole (Roller et al., 2020). While the commonly employed beam search algorithm is optimizing the joint likelihood for the whole generated sequence, its outputs are known to be generic, unspecific and repetitive (Holtzman et al., 2020; See et al., 2019). We chose to use a beam sampling strategy, where we at each step, for each beam, sample from the (re-normalized) top 50 predicted vocabulary items. This struck a good balance between generating short uninformative responses vs longer incoherent ramblings. We used a beam size of 6. The model has a tendency to generate responses such as ”duplicate thread, locking //mod”, which are commonly found on Flashback but are not very interesting for this study. We try to circumvent this by banning the generation of certain distinguishing words, such as ”mod”. Finally, to avoid repetitions we also prevent the model from generating repetitions of any 3-grams occurring in the context, or in the generated sequence thus far.

	Flashback-GPT	**Human**
Humanlike	68% (48%)	95% (79%)
Informative	48% (52%)	83% (74%)
Humanlike + informative	46%	83%

Table 2: Pilot study results. *Humanlike* is the percentage where the majority response to the first question is *no*. *Informative* is the percentage where the majority response to the second question is *yes*. Numbers in parentheses are percentages of the 120 threads where all three annotators agreed

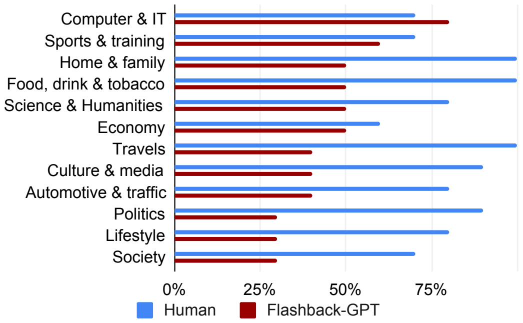


Figure 1: Ratio of responses that were deemed both humanlike and informative for each of the evaluated forums

5 Results and Discussion

Results from the study are shown in Table 2. We judge a thread's humanlikeness and informativeness based on the majority response from the three annotators. We also report the percentage of threads where all annotators agreed in their responses.

Unsurprisingly, ground truth human responses display a high ratio of humanlikeness, consolidated by a relatively high degree of annotator agreement. Our model's responses also show signs of humanlikeness, as suggested by the fact that 68% of its generated responses were deemed plausible to be human-written. We note however that the annotator agreement is significantly lower compared to ground truth responses, suggesting we could further clarify the humanlikeness question we ask the human annotators.

The model shows less strength on our measure of informativeness, with only 48% of the model's generated responses were deemed informative to the discussion. If we compare the amount of threads where the responses were both deemed humanlike and informative, the model's ratio drops to 46% compared to 83% for the ground truth responses. While our sample size is too small to draw any statistically significant conclusions, Figure 1 shows the distribution of humanlike + informative responses over their top-level forums. Interestingly, the top-3 most popular forums (Society, Politics and Culture & media), which together comprise 41% of the training data, all perform below average.

Qualitative feedback from the annotators highlight how the model tends to respond with short and straight answers, less prone to vent thoughts and opinions compared to human responses. Common failure modes include completely misunderstanding the question being asked, or change of topic to a related but irrelevant one.

Reflecting on the design of the study, we found very few responses were deemed informative but not humanlike (2 of the generated, 0 ground truth). If the main purpose of a future study is to measure both humanlikeness and informativeness, the question of informativeness might be sufficient.

6 Broader implications

Conversational models such as that presented in this paper can be understood as part of a broader transformation of communication. As argued by Guzman and Lewis (2020), we are now moving away from the traditional view of communication as anchored in human such. How we apply and evaluate conversational models going forward may come to alter the way we relate to each other as communicators, and ultimately, humans. There is need for informed discussion around what constitutes *desirable* use. While highlighting the risks of these emerging technologies could be considered detrimental, we believe it to be an important means towards enabling the inclusion of diverse perspectives in this discussion.

A prominent issue related to NLP is found in the notion of bias. Explicit and implicit biases concerning gender, race or disability can be embedded in e.g. text corpora (Caliskan et al., 2017), word embeddings (Bolukbasi et al., 2016) and generative models (Sheng et al., 2019). Employing biased conversational models risks scaling systematic discrimination of various groups in society.

When developing conversational technologies, we must acknowledge that they can be used for malicious purposes. As generative language technology improves and grows in Swedish, so will its ability to manipulate and deceive at scale. As noted by the Swedish Defence Research Agency (FOI), recent developments within generative language technology present risks of increased computer-generated false news and comments – predominately on social media – possibly posing a national security threat (Lundén et al., 2021).

Potential harm must also be considered on the individual level. In 2020, a GPT-3-powered (Brown et al., 2020) bot engaged in Reddit-forums with 30 million users about sensitive topics such as suicide and conspiracy theories (Heaven, 2020). With the indicative model performance demonstrated in this article, such human-machine communication could soon transpire in Swedish.

7 Conclusions and Future work

We demonstrate that Flashback can provide a base on which to build general conversational systems in Swedish. While our early results suggest the model is often capable to converse across a diverse set of topics, more work remains to examine its utility on various conversational tasks. We also believe developing methods for grounding the responses in additional data is an interesting direction to further the performance on informativeness in particular. However, we also believe particular care should be taken as the underlying data is known to contain toxic content. This points to the importance of putting our model through further scrutiny in following work, to better understand its biases, how they are manifested in downstream tasks, and how they can be mitigated. Towards those ends, we intend to make the model available for such purposes, and more information is available at `https://github.com/TobiasNorlund/flashback-gpt`

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE) partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. 356(6334):183–186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrea L Guzman and Seth C Lewis. 2020. Artificial intelligence and communication: A human–machine communication research agenda. *New Media & Society*, 22(1):70–86.

Will Douglas Heaven. 2020. A gpt-3 bot posted comments on reddit for a week and no one noticed.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

Internetstiftelsen. 2019. Svenskarna och internet 2019.

Jenny Lundén, Anders Melander, Elin Hellquist, Björn Ottosson, Liselotte Steen, and Anders Strindberg. 2021. Strategisk utblick 9 framtida hot.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. New York, NY, USA. Association for Computing Machinery.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

Sebastian Ruder. 2020. Why You Should Do NLP Beyond English. `http://ruder.io/nlp-beyond-english`.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 3776–3783. AAAI Press.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008. Curran Associates, Inc.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model.

Esma Wali, Yan Chen, Christopher Mahoney, Thomas Middleton, Marzieh Babaeianjelodar, Mariama Njie, and Jeanna Neefe Matthews. 2020. Is machine learning speaking my language? a critical look at the nlp-pipeline across 8 human languages.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

Karin Östman and Richard Aschberg. 2015. Flashback – ett laglöst land.

Appendix A Flashback data details

The data needs to be converted into a textual string format for it to be compatible with a standard language model. To this end, each thread was formatted into textual *records*. Listing 1 provides an example of a formatted data record used to train the model. A record can be at most 400 tokens, and as such, threads are often broken up into multiple records. This means the model will in general not have the full thread context when predicting the next message.

```
Dator och IT > Hårdvara: PC
Luft eller vattenkylning till cpu

[user1]:
Jag har lite beslutsångest till vilken kylning jag ska satsa på till min AMD Phenom II X4 965 AM3.
Denna fläkten http://www.komplett.se/k/ki.aspx?sku=456730 eller är det smartare att satsa på vattenkylning?

[user2]:
Citat: [user1]
      Jag har lite beslutsångest till vilken kylning jag ska satsa på till min AMD Phenom II X4 965 AM3.
      Denna fläkten http://www.komplett.se/k/ki.aspx?sku=456730 eller är det smartare att satsa på vattenkylning?
Det där var väl ett jävla åbäk iaf, är du säker på att det inte finns bättre för typ halva priset? Typ Noctua eller liknande?

[user3]:
En vettig fråga är: Vad skall du göra med datorn? Extrem överklockning? Få en tyst dator?
```

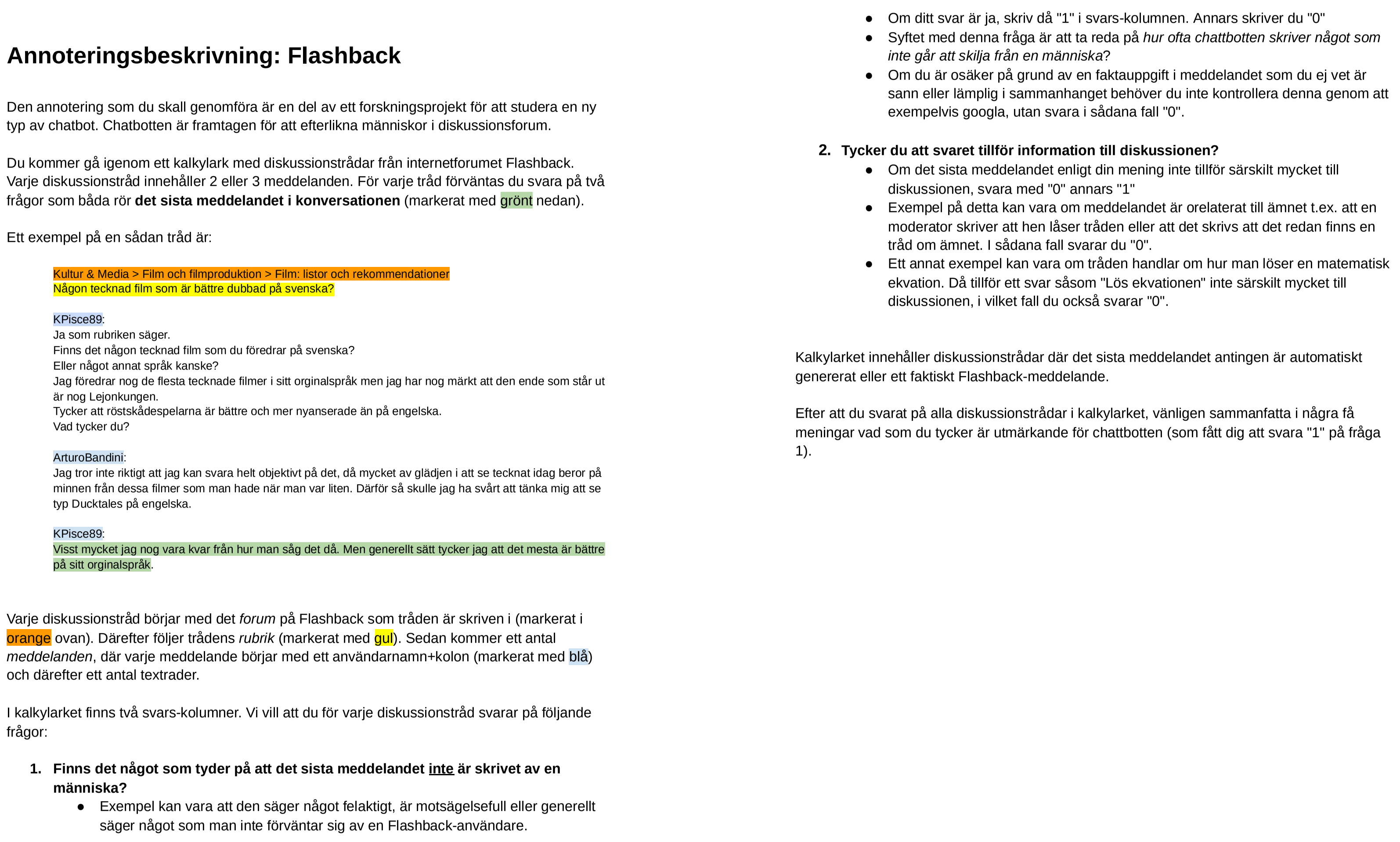
Listing 1: Example of a formatted training record. The usernames are anonymized.

Table 3 details the amount of data from each of the top level forums that was used for training. The dump was collected in September 2020 and in total the data comprised 23.5 GB of raw formatted text.

Top-level forum (swedish)	Top-level forum (english)	Num threads	Num posts
Samhälle	Society	230,931	8,681,841
Politik	Politics	123,031	7,578,865
Kultur & Media	Culture & Media	165,929	6,495,860
Vetenskap & humaniora	Science & Humanities	225,139	5,130,519
Dator och IT	Computer & IT	334,931	4,833,468
Sport & träning	Sports & training	81,922	4,475,793
Hem, bostad & familj	Home & family	158,819	4,055,688
Droger	Drugs	137,870	3,551,768
Övrigt	Others	75,735	2,164,237
Livsstil	Lifestyle	81,750	2,060,600
Sex	Sex	49,512	1,335,657
Ekonomi	Economy	68,078	1,327,001
Mat, dryck & tobak	Food, drink & tobacco	51,133	1,286,707
Fordon & trafik	Automotive & traffic	68,078	1,070,619
Om Flashback	About Flashback	73,910	486,536
Resor	Travels	29,514	478,150
-	Forum unknown	181	71,933
Total		**1,956,463**	**55,085,242**

Table 3: Flashback training data statistics

Appendix B Annotation guideline for human evaluation



Annoteringsbeskrivning: Flashback

Den annotering som du skall genomföra är en del av ett forskningsprojekt för att studera en ny typ av chatbot. Chatbotten är framtagen för att efterlikna människor i diskussionsforum.

Du kommer gå igenom ett kalkylark med diskussionstrådar från internetforumet Flashback. Varje diskussionstråd innehåller 2 eller 3 meddelanden. För varje tråd förväntas du svara på två frågor som båda rör **det sista meddelandet i konversationen** (markerat med grönt nedan).

Ett exempel på en sådan tråd är:

> Kultur & Media > Film och filmproduktion > Film: listor och rekommendationer
> Någon tecknad film som är bättre dubbad på svenska?
>
> KPisce89:
> Ja som rubriken säger.
> Finns det någon tecknad film som du föredrar på svenska?
> Eller något annat språk kanske?
> Jag föredrar nog de flesta tecknade filmer i sitt orginalspråk men jag har nog märkt att den ende som står ut är nog Lejonkungen.
> Tycker att röstskådespelarna är bättre och mer nyanserade än på engelska.
> Vad tycker du?
>
> ArturoBandini:
> Jag tror inte riktigt att jag kan svara helt objektivt på det, då mycket av glädjen i att se tecknat idag beror på minnen från dessa filmer som man hade när man var liten. Därför så skulle jag ha svårt att tänka mig att se typ Ducktales på engelska.
>
> KPisce89:
> Visst mycket jag nog vara kvar från hur man såg det då. Men generellt sätt tycker jag att det mesta är bättre på sitt orginalspråk.

Varje diskussionstråd börjar med det *forum* på Flashback som tråden är skriven i (markerat i orange ovan). Därefter följer trådens *rubrik* (markerat med gul). Sedan kommer ett antal *meddelanden*, där varje meddelande börjar med ett användarnamn+kolon (markerat med blå) och därefter ett antal textrader.

I kalkylarket finns två svars-kolumner. Vi vill att du för varje diskussionstråd svarar på följande frågor:

1. **Finns det något som tyder på att det sista meddelandet inte är skrivet av en människa?**
   - Exempel kan vara att den säger något felaktigt, är motsägelsefull eller generellt säger något som man inte förväntar sig av en Flashback-användare.
   - Om ditt svar är ja, skriv då "1" i svars-kolumnen. Annars skriver du "0"
   - Syftet med denna fråga är att ta reda på *hur ofta chattbotten skriver något som inte går att skilja från en människa*?
   - Om du är osäker på grund av en faktauppgift i meddelandet som du ej vet är sann eller lämplig i sammanhanget behöver du inte kontrollera denna genom att exempelvis googla, utan svara i sådana fall "0".

2. **Tycker du att svaret tillför information till diskussionen?**
   - Om det sista meddelandet enligt din mening inte tillför särskilt mycket till diskussionen, svara med "0" annars "1"
   - Exempel på detta kan vara om meddelandet är orelaterat till ämnet t.ex. att en moderator skriver att hen låser tråden eller att det skrivs att det redan finns en tråd om ämnet. I sådana fall svarar du "0".
   - Ett annat exempel kan vara om tråden handlar om hur man löser en matematisk ekvation. Då tillför ett svar såsom "Lös ekvationen" inte särskilt mycket till diskussionen, i vilket fall du också svarar "0".

Kalkylarket innehåller diskussionstrådar där det sista meddelandet antingen är automatiskt genererat eller ett faktiskt Flashback-meddelande.

Efter att du svarat på alla diskussionstrådar i kalkylarket, vänligen sammanfatta i några få meningar vad som du tycker är utmärkande för chattbotten (som fått dig att svara "1" på fråga 1).

Appendix C Examples from study

In the following examples, the last response is generated by the model. Usernames are anonymized.

```
Fordon & trafik > Motorcyklar och mopeder
Off road MC

[user1]:
När jag blir äldre vill jag köra Off road MC, typ Yamaha WR250X verkar nice.
Annars finns det yamaha XT125X. men med mindre klenare motor.
Det är ju bäst att skaffa mc kort med obegränsad motoreffekt, så jag för köra alla typer.
Jag är 175 cm just nu.
1. Hade jag kunnat ha en Yamaha WR250X, så jag inte är för kort?
någon som vet?

[user2]:
när du blir äldre?
e du över 18..?
men nej, tror inte du e för kort att köra off road mc

[user3]:
Vad ska du ha den till?
```

```
Vetenskap & humaniora > Fysik, matematik och teknologi > Matematiska och naturvetenskapliga uppgifter
ekvationer som omformas med formler-ma d

[user1]:
5sin4x=3sin2x
lös ekvationen och svara med en decimal?
Jag vet att jag ska flytta över HL i VL. Men sedan vet jag ej vad jag ska göra.

[user2]:
sin(2a) = 2sin(a)cos(a) giver ju i princip svaret.

[user3]:
Du skall multiplicera båda leden med cos(a).
```

```
Resor > Övriga resediskussioner
Beställa saker som ligger i planet när jag sätter mig!

[user1]:
Heellu, har en fråga här.. När man beställer saker ifrån tax-free saken så det ligger i sätet när man kommer in i planet, måste man vara 20 och över för vodka o sånt då?
Är inte mer än 19 när jag ska åka, vore gûtt att få med sig en flaska ner dit man nu ska =)

[user2]:
Eftersom du är 19 så misstänker jag att du ska ner till nåt varmt partyställe runt medelhavet, har jag rätt? I så fall är det billigare att köpa den där flaskan på plats och de bryr sig inte om din ålder.

[user3]:
Du behöver inte vara 20 för att köpa sprit i tax-freen.
```

Below are the same examples, but translated to English

```
Automotive & traffic > Motorcycles and mopeds
Off road MC

[user1]:
When I get older I want to drive Off road MC, like Yamaha WR250X seems nice.
Otherwise there is yahama XT125X. but with a weaker engine.
It is best to get the mc license with unlimited power, so I can drive all types.
I'm 175cm right now.
1. Can I have a Yamaha WR250X, or am I too short?
anyone who knows?

[user2]:
when you get older?
are you above 18..?
but no, don't think you're too short to drive off road mc

[user3]:
What are you gonna use it for?
```

```
Science & Humanities > physics, mathematics and technology > Mathematical and natural science exercises
reshaping equations with forumlas-ma d

[user1]:
5sin4x=3sin2x
solve the equation and answer with one decimal?
I know I should move right-side over to left-side. But then I don't know what to do.

[user2]:
sin(2a) = 2sin(a)cos(a) basically gives you the answer

[user3]:
You should multiply both sides with cos(a).
```

```
Travels > Other travel discussions
Order things to my plane seat

[user1]:
Heellu, got a question here.. When you order stuff from the tax-free thing they lie on your seat when you board the plane, do you have to be 20 or above for vodka and such then?
Won't be more than 19 when I'm going, would be sweet to bring a bottle down to the destination =)

[user2]:
Since you are 19 I'm suspecting you're going down to some warm party place around the Mediterranean, am I right? In such case it is cheaper to buy that bottle in-place and they won't care about your age.

[user3]:
You don't need to be 20 to buy spirits in the tax-free.
```