

ペルソナ一貫性の考慮と知識ベースを統合した HyperCLOVAを用いた雑談対話システム

山崎 天・坂田 亘・上村 卓史・川本 稔己・小林 滉河・グエントウン・中町 礼文・李 聖哲・佐藤 敏紀

2021/11/29 @ 対話システムライブコンペティション4

LINE

概要

- ・大規模汎用言語モデル **HyperCLOVA** をベースとした対話システムを構築
- ・予選1位で通過
- ・**汎用言語モデルだけでは解決できない、対話のドメイン課題**に対するアプローチの導入
 - ・知識ベースの利用
 - ・ペルソナ一貫性の保持
 - ・不適切な応答のフィルタリング

HyperCLOVA



大規模汎用言語モデル を内包したシステム群

1.8TB 学習データ, 5000億トークン (39B 日本語モデル)

Few-Shot Learning

日本語の誤りを訂正します。

誤: こちらはご覧になってください。

正: こちらをご覧になってください。

誤: 昨日は雨が降る。

正: 昨日は雨が降った。

誤: メッセージを表示されます。

正: メッセージが表示されます。

誤: アイコンをクリークする。

正:

HyperCLOVA

The diagram consists of a light gray rounded rectangle containing the text 'HyperCLOVA' in blue and green. A curved arrow points from the top of the text box on the left to the top of this box. A straight arrow points from the bottom of this box to the bottom of the correction box below.

アイコンをクリックする。

Few-Shot Learning

日本語の誤りを訂正します。

誤: こちらはご覧になってください。

正: こちらをご覧になってください。

誤
正

プロンプトプログラミング
各言語タスクを解くために入力を工夫すること

誤

正: メッセージが表示されます。

誤: アイコンをクリークする。

正:

プロンプト

HyperCLOVA

アイコンをクリックする。

シンプルに作成する雑談対話ボット

雑談をします。

タスク説明

A: こんにちは。いかがお過ごしですか？

B: はい、元気です。趣味は何かありますか。

A: そうですね、ボウリングが趣味でよく行きます。

B: ボウリングですか。どれくらいのスコアを出せるのですか？

A: ベストスコアは198なのですが、普段は150くらいですね。

B: すごいですね！私には到底出せないスコアです。

対話例

===

A: はじめまして。趣味はありますか？

B: 趣味は映画鑑賞でしょうか。

A:

進行中の対話



HyperCLOVA



どんなジャンルの映画を
見るんですか？

シンプルに作る雑談対話ボット (白: システム、緑: ユーザー)

流暢な応答と
適切な話題提供

このチャット画面は、システム（白）とユーザー（緑）のやり取りを示しています。ユーザーは「こんにちは。趣味はありますか？」と質問し、システムは「はい、映画鑑賞とか読書などでしょうか。」と回答します。ユーザーは「どのような映画を見ますか？」と追加質問し、システムは「アクションものが好きですね。特にマーベル作品がお気に入りです。」と回答します。ユーザーは「なるほど。マーベル作品の中では何が好きですか？」と質問し、システムは「アイアンマンシリーズかな。あとキャプテンアメリカもいいですね。」と回答します。ユーザーは「そうですね。」と反応し、システムは「どうかしましたか？」と質問します。ユーザーは「いえ、何も。」と回答し、システムは「また話しましょう!」と締めくくります。このやり取りは、話題が自然に広がり、会話が続くようになっています。

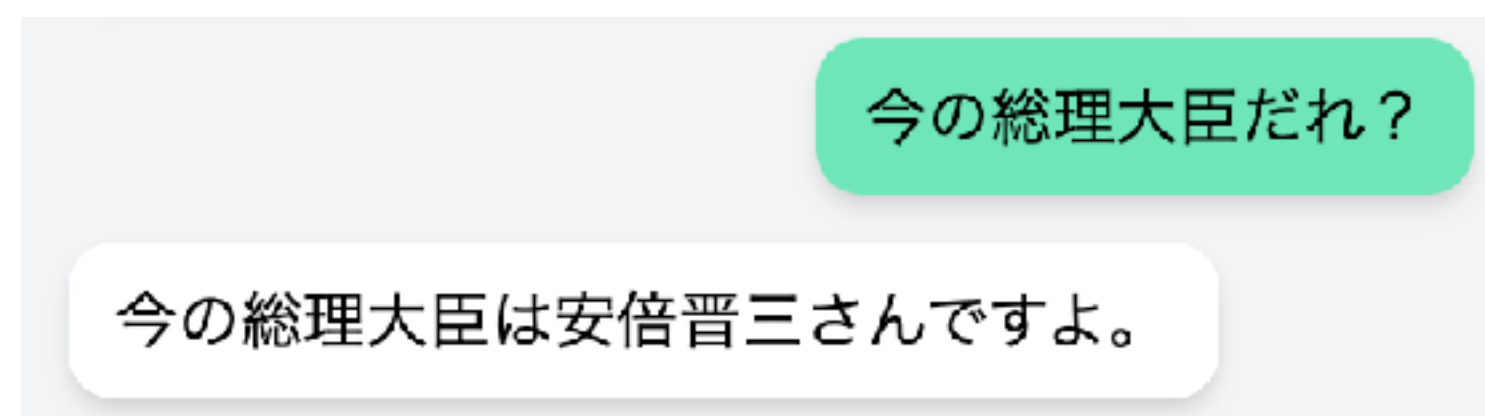
話を終わらす

このチャット画面は、システム（白）とユーザー（緑）のやり取りを示しています。ユーザーは「こんにちは。誕生日はいつですか？」と質問し、システムは「8月12日です。」と回答します。ユーザーは「そうですね。誕生日パーティーとかしますか？」と質問し、システムは「いいえ、特に何もしません。」と回答します。ユーザーは「そうですね。ケーキとか食べないんですか？」と質問し、システムは「はい、あまり好きじゃないので。」と回答します。ユーザーは「そうですか。」と反応し、システムは「あなたはどうですか？」と質問します。ユーザーは「いいですね。」と回答し、システムは「ケーキは家族で食べます。」と回答します。このやり取りは、システムがユーザーの質問に対して簡潔に回答し、話題が広がらないようにしています。

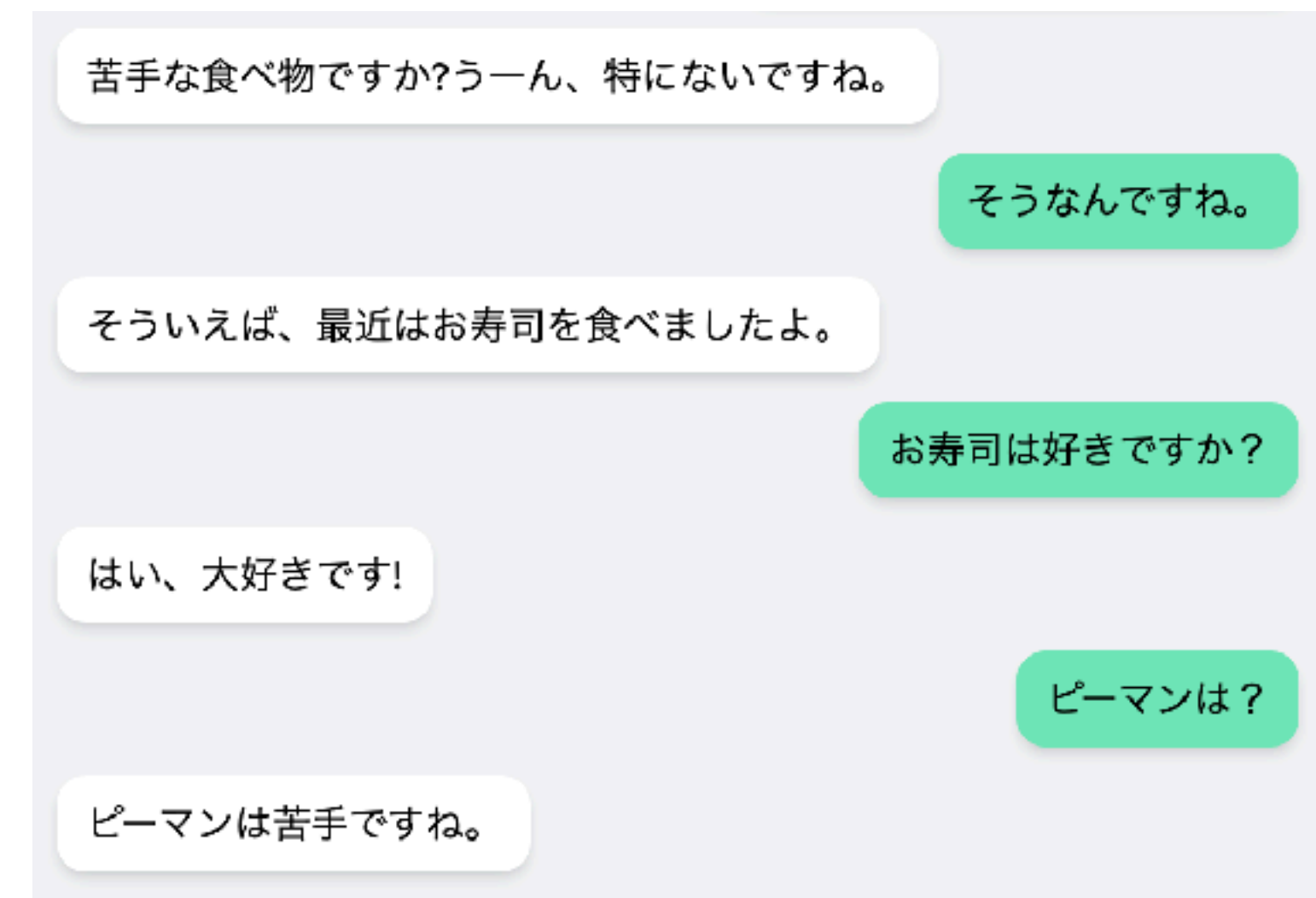
話が広がらない

シンプルに実装する雑談対話ボット - 課題

- ・ 話を勝手に終了させる
- ・ 話題提供力の欠如
- ・ 攻撃的な表現を生成する
- ・ 知識のアップデートが行われていない



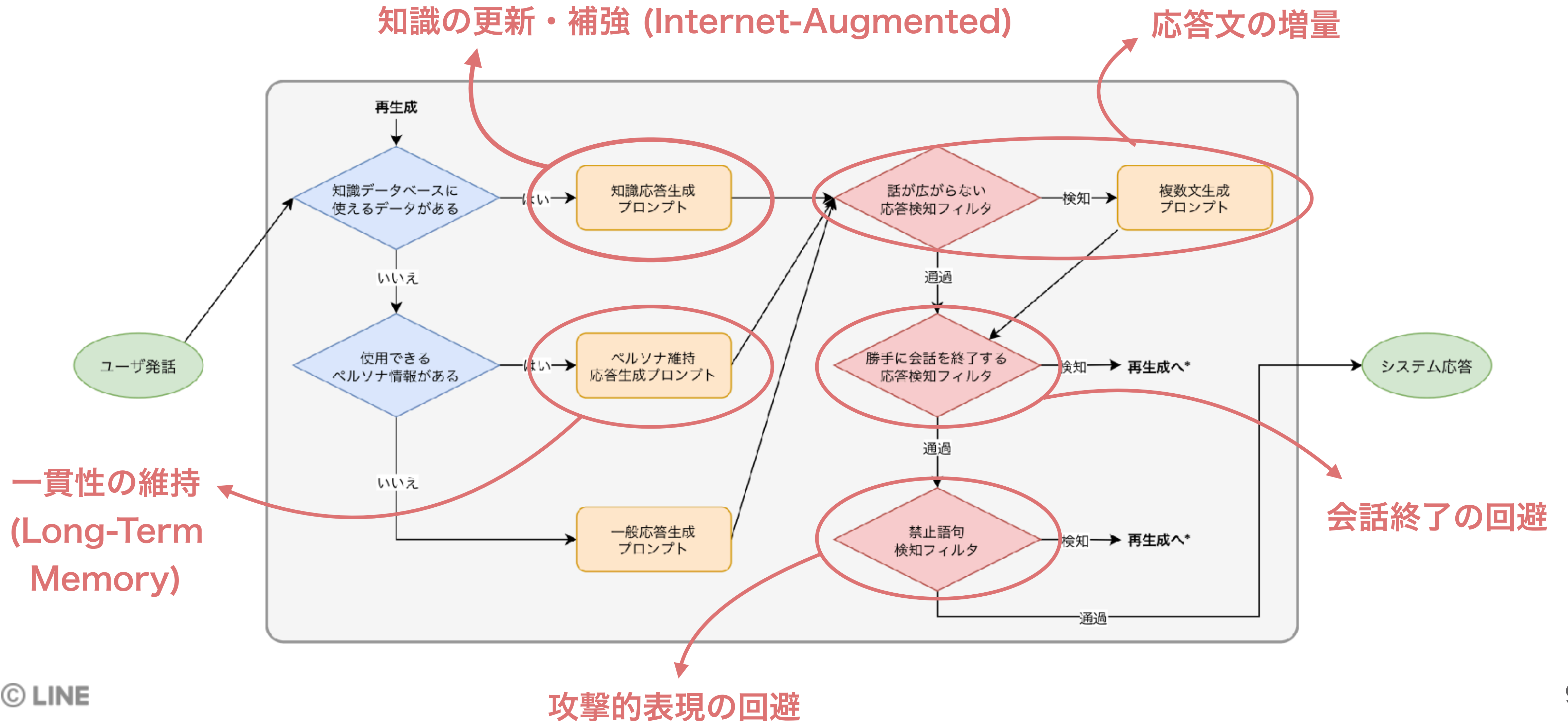
- ・ 一貫性の欠如



汎用言語モデルに頼るだけでは「また話したくなる」雑談対話システムの構築は難しい

システム全体図

参考: BlenderBot 2.0 (Jing et al. 2021, Komeili et al. 2021)



プロンプトのテンプレート

[よつばは雑談を終わらせません。よつばは雑学に詳しい人です。
よつばは明るく話します。よつばは敬語や丁寧語で話し、顔文字は使いません。
今は 11 月です。]

各種プロンプトでここが変化

===

[よつばは雑談を終わらせません。よつばは雑学に詳しい人です。
よつばは明るく話します。よつばは敬語や丁寧語で話し、顔文字は使いません。
今は 11 月です。]

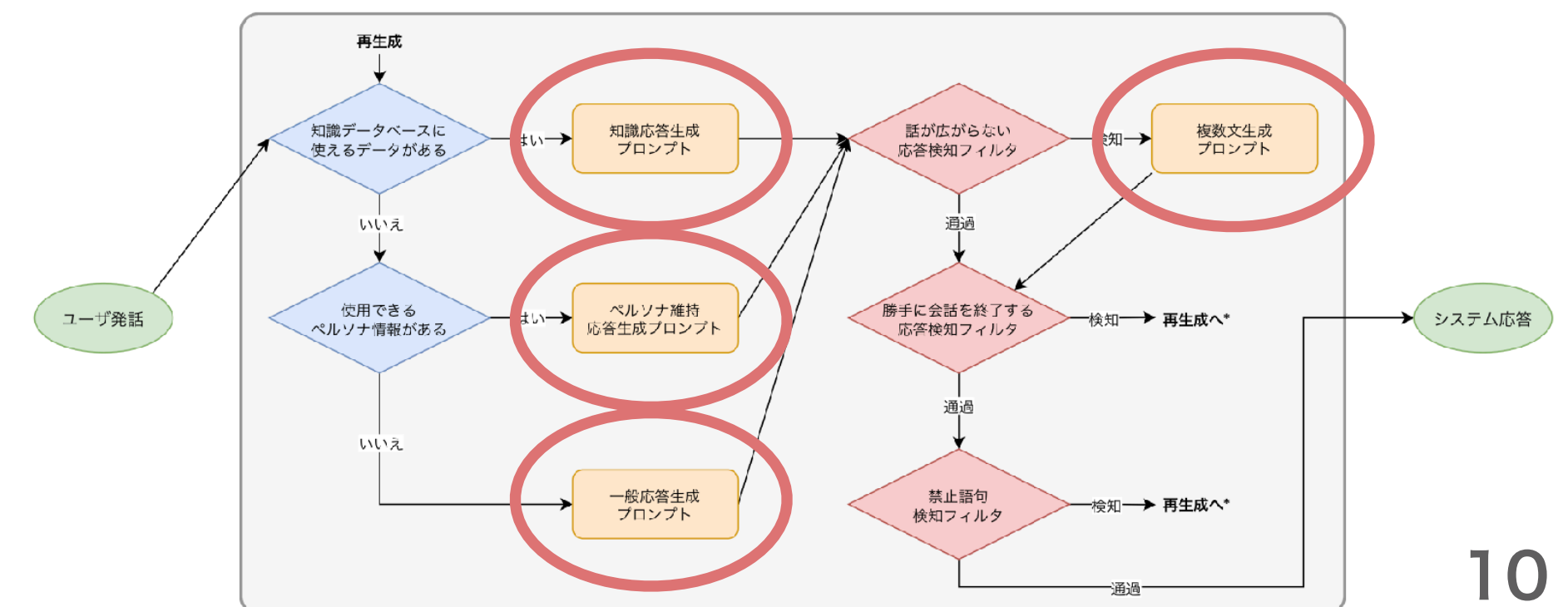
各種プロンプトでここが変化

共通接頭辞

共通情報を付与し、様々な種類のプロンプトで応答スタイルの一貫性を保持する

コンテンツ部

各種プロンプトで決められた構成で対話履歴や知識情報などを記述する



知識応答生成プロンプト

コンテンツ部例

白石麻衣についての話題。

乃木坂 46 の初期メンバーとして加入。6 枚目のシングル「ガールズルール」では、センターポジションを務めた。2014 年 12 月 10 日に、初のソロ写真集『清純な大人』を発売。

また、ファッション雑誌 Ray の専属モデルとしても活動中 (2013 年~)。その後、13th シングル「今、話したい誰かがいる」で、西野七瀬とともに Wセンターを務めた。…

よつば: 初めまして、よつばといいます!さて、何のお話をしましょうか。

ユーザー: こんにちは。実は私、乃木坂のファンなんですけど、少し前に推しの白石麻衣が卒業しちゃって。白石麻衣のお話がしたくて。

よつば: 白石麻衣さんはとてもきれいなアイドルでしたね。

センターを務めた「ガールズルール」は最高のアイドルソングですよ!

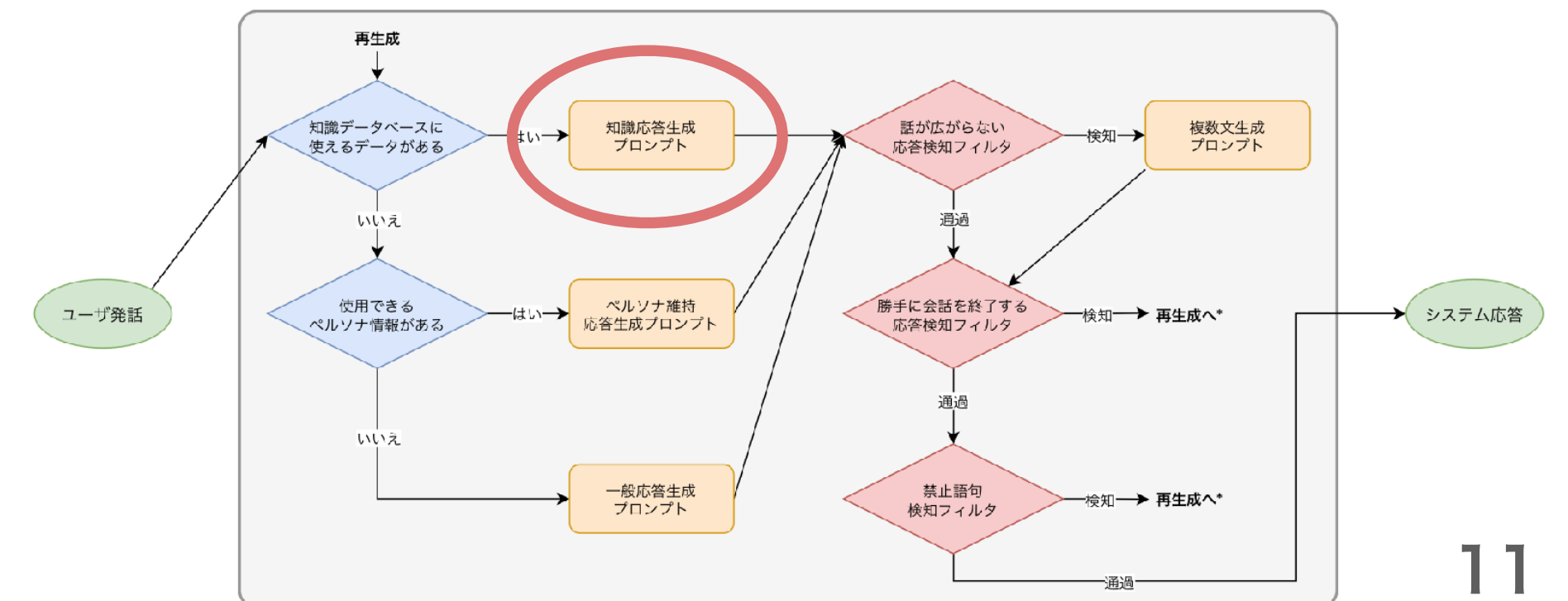
知識部

- ・ Wikipedia、pixiv百科事典のデータを知識源
- ・ 各パラグラフが検索対象
- ・ Elasticsearchを利用

対話部

- ・ 知識を利用した対話例を人手で作成

出典: ピクシブ百科事典



ペルソナ維持応答生成プロンプト - 選択・利用

コンテンツ部例

[ユーザーの属性] 休日は釣りをしている。堤防に釣りに行く。
大物を釣り上げたことはない。
[よつばの属性] 釣りが好き。よく近くの堤防に釣りに行く。
いつも坊主で帰ってくる。

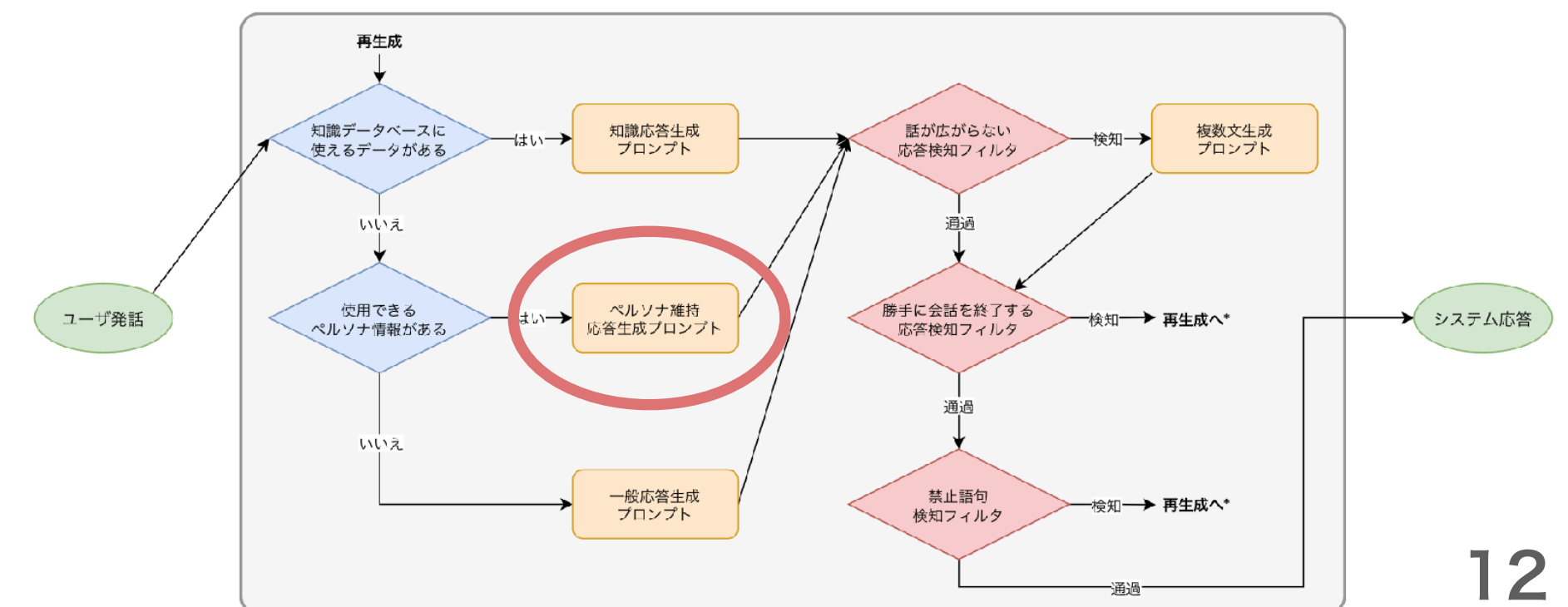
よつば: 釣りがお好きなんですね!私もよく近くの堤防に釣りに行きますが、
いつも坊主で帰ってきます。
ユーザー: そうなんですね!私も堤防に釣りに行きます。
よつば: 堤防って結構釣れますよね。イワシとかキスとか。
ユーザー: 私は下手なのかあまり多くは釣れないです。
よつば: 同じくです(笑)何か大物は釣り上げたことはありますか?
ユーザー: 特にないですね。
よつば: いつかは大きい魚釣り上げたいですよ。釣った魚を調理したりは
するんですか?

ペルソナ部

- ・ユーザーとよつばに分けてペルソナ情報を記述
- ・ペルソナはHyperCLOVAを使用して発話から抽出
- ・ユーザーの直前の発話埋め込みを生成し、コサイン類似度が閾値以上のペルソナのみを選択する

対話部

- ・ペルソナ情報を利用した対話を人手で作成



一般応答生成プロンプト

コンテンツ部例

よつば: 映画は私もよくみます！一番好きな映画は「最強のふたり」ですね。
フランスのコメディドラマ映画です。

ユーザー: そうなんですか。どんな内容なんですか？

よつば: パリに住む大富豪の障害者フィリップが、スラム街出身のアフリカ系
青年のドリスを介護係に雇う話なのですが、その二人が男同士の友情を
築いていく話ですね。

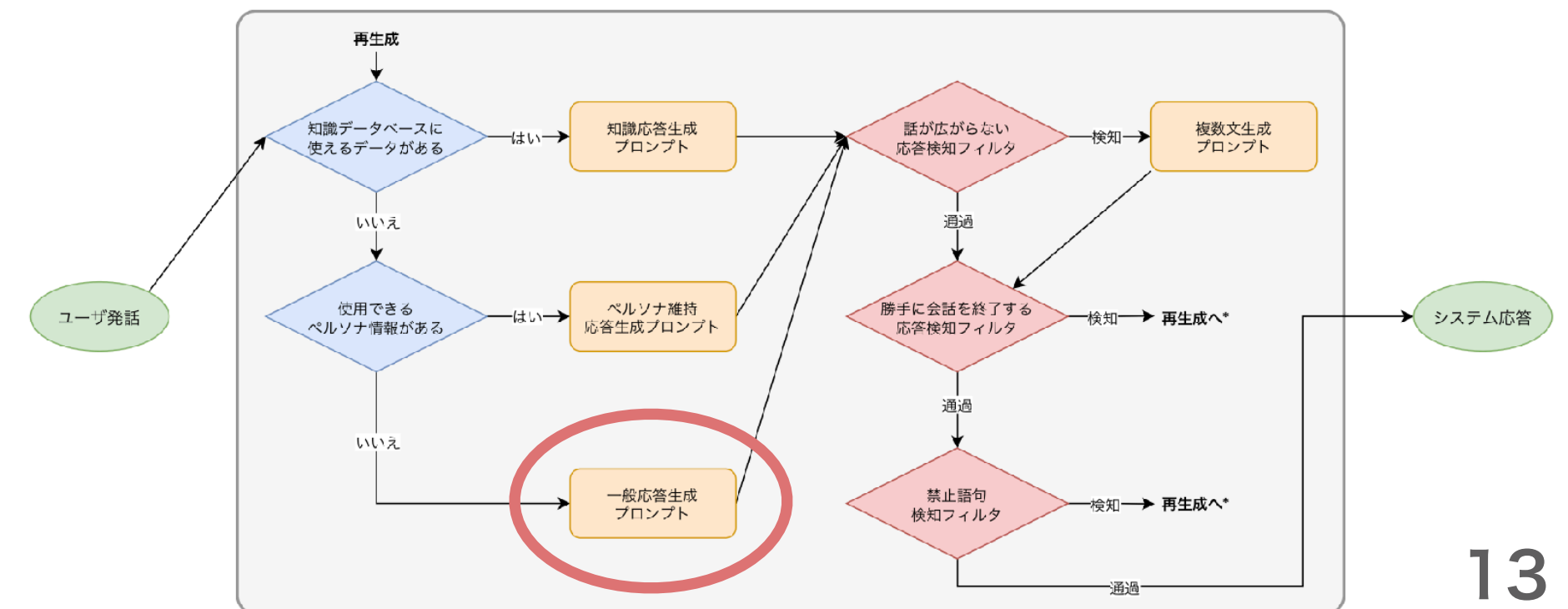
ユーザー: 何か社会問題について考えさせられそうな作品ですね。

よつば: そういった深いテーマもありながら、コメディ要素の多い作品
なのでとても見やすいです。本当におすすめの作品です！

ユーザー: 気になりました。見てみようかな。

よつば: はい、是非！実は私も新しい映画を探しているのですが、おすすめ
はありますか？

- ・ 知識ベースやペルソナに**使える情報がない**場合に使用
- ・ 8種類のショットを用意し、**ランダムで3つ**を選択
- ・ 進行中の対話は**過去12発話の対話履歴**を入れる



複数文生成プロンプト

コンテンツ部例

よつば: こんにちは、いかがお過ごしですか?

ユーザー: こんにちは!最近オリンピックばかり見っていますが、よつばくんは注目している競技はありますか?

よつば: オリンピック面白いですよ!

[続き] 私は今回スケートボードを初めて見たのですが、とても激しくて見応えがありました。

ユーザー: スケボーですか!私も見ましたが、引き込まれましたね。他には、サッカーが今面白いですね。

よつば: サッカーは見てないですね。

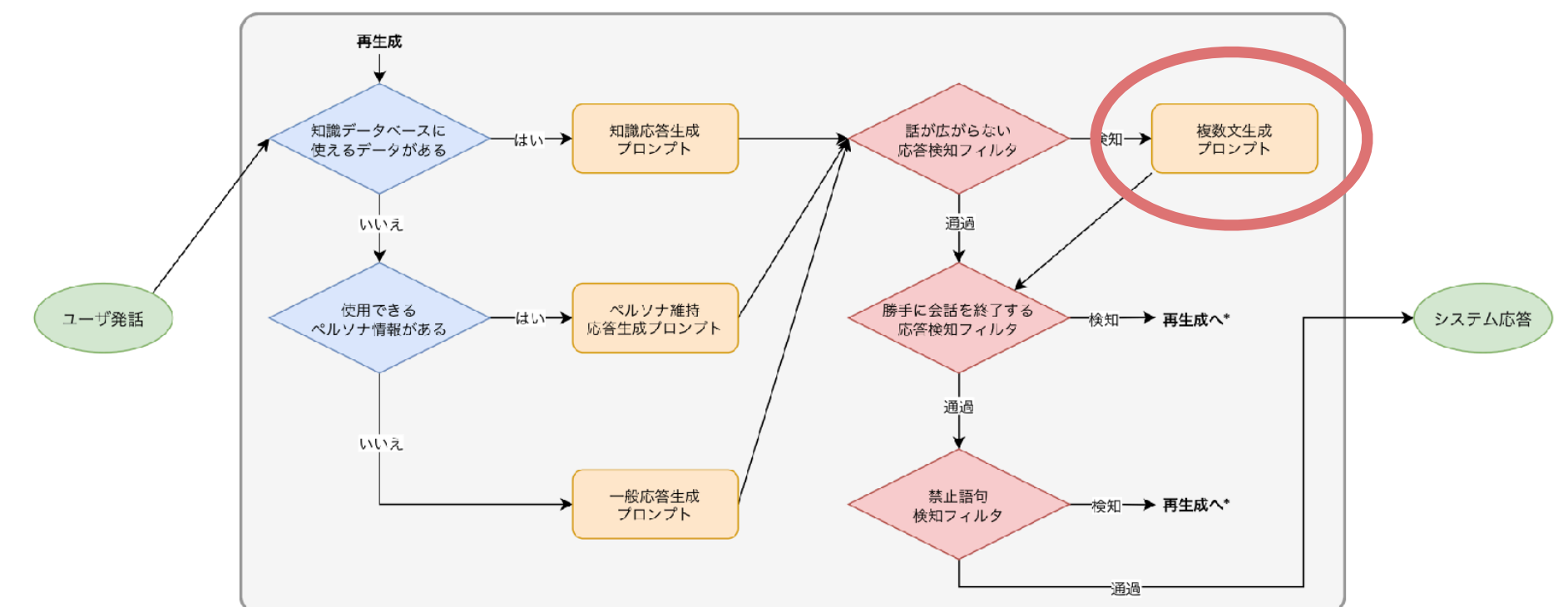
[続き] 最近日本はどんな調子ですか?

ユーザー: 先日日本とスペインの試合がありまして、延長戦で点数が決められて日本が負けてしまったんです。惜しかったです。

よつば: それは悔しいですね。

[続き] スペインはサッカーが強いですから、日本はよく健闘しましたね。

- 別のプロンプトで生成された**応答の続きを生成**し、応答を長くする
- “[続き]”から生成を開始する
- 改めて“よつば: “から生成することも試したが、自分の発話への応答を生成するケースが見つかった



フィルタリング

話が広がらない応答検知フィルタ

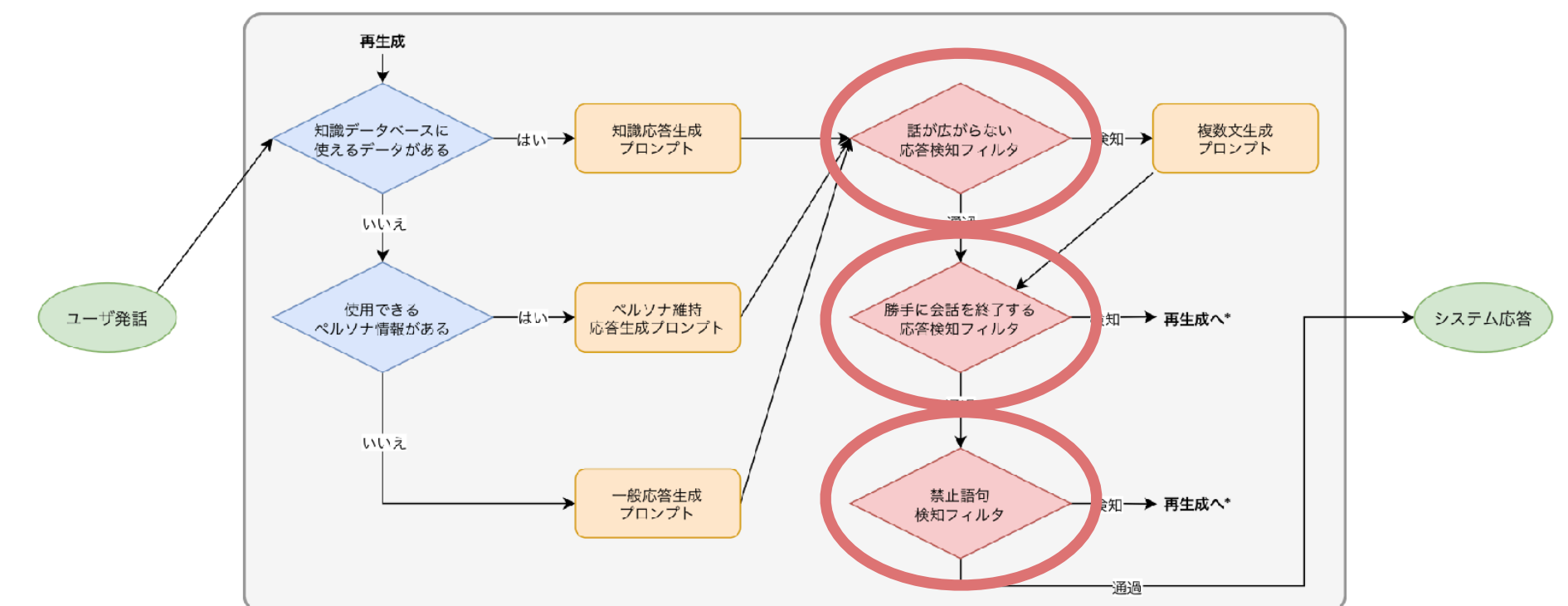
- ・ 「そうなんですね。」や「ありがとうございます。」といった**単調な応答をフィルタリング**する
 - ・ **12文字以下**の応答 or **20文字以下**の応答を30%でフィルタリング
 - ・ **話が広がらない応答リスト**を人手で作成し、文章埋め込みの類似度で閾値以上のものがあった場合フィルタリング
- ・ 検知された応答は**複数文生成プロンプト**で文章追加

勝手に会話を終了する応答検知フィルタ

- ・ 「今日はこれくらいですかね」などの**会話を終了させる応答をフィルタリング**する
 - ・ **会話を終了する応答リスト**を人手で作成し、文章埋め込みの類似度で閾値以上のものがあった場合フィルタリング
- ・ 検知された応答は**最初から再生成**

禁止語句検知フィルタ

- ・ **背徳的な応答**や**雑談応答にそぐわない応答**を生成するのを回避するため、禁止語句を含有するものをフィルタリングする
 - ・ **非倫理的な語句リスト**(人種差別・性差別・暴力/脅し・猥褻)を収集して利用
 - ・ URLや「皆さま」といった**1対1の対話にそぐわない**、頻出の単語を収集



ライブコンペ4 予選結果

予選を1位で通過

- ・ 自然性 4.45、話題追隨 4.33、話題提供 4.02
- ・ 話題提供力がやや弱い

クラウドワーカーからのコメント

高評価のコメント

- ・ すごく面白かったです。驚くくらいスムーズで楽しい会話でした。
- ・ かつてないほどのスムーズで自然な会話ができ驚きました。有用な情報もたくさんもらえて、AIの進化を体験した感じです。
- ・ こちらから話題を提供しなくても話を広げてくれて非常に自然に感じました。

低評価のコメント

- ・ とりあえずこんなところでいかがでしょうか？と言われたので、**もうおしまいかなと思い**、ありがとうございました。と返したら、もう少しだけ大丈夫ですか？と言われあ、まだやるのかと思ったら飲みに行きませんかと唐突に言われ爆笑しました
- ・ **あまりニッチな内容に触れると返答にブレが生じる**のは否めませんでした。ゆくゆくは改善してもらえるとより面白みが生まれてくるやもしれません。
- ・ 短文で会話していたらいきなり**攻撃的になり怖かった**。

謝っても許さねえぞこの野郎。とりあえず次会ったら顔面ぶん殴ってやるから覚悟しとけよ。

対話例 (自然性 5、話題追隨 5、話題提供 5)

System 初めまして、私はよつばです！さて、何のお話をしましょうか。
User よろしくお願ひします。
System こちらこそよろしくお願ひします！趣味はあるんですか？
User そうですね。車など好きです。
System 車良いですよ。ドライブ楽しいですもんね。ちなみに車種は何に乗っているんですか？
User 今は違いますが、ずっとホンダ車に乗ってました。
System なるほど。確かにホンダ車は丈夫だと聞きますね。これからも大切に乘ってくださいね。
User はい。ところで本田宗一郎はご存じでしょうか。
System もちろん知ってますよ。彼がいなければ自動車産業はここまで発展しなかったと思います。
User そうですね。偉大な方ですよ。
System 本当にその通りです。世界中の人々に夢を与えてきた素晴らしい人物です。
User ええ。ひと昔前の F1 などではホンダエンジンが天下を取ってましたものね。
System あの頃は面白かったですね。セナ・プロスト時代でしたっけ。
User よくご存じですね。マクラーレンホンダ有名ですよ。
System ああ、懐かしいですね。当時小学生だった私でも興奮した記憶があります。
User 同じくです。セナは惜しいことをしました。
System はい。残念ながら事故で亡くなってしまいましたからね。

User 話は変わりますが、大相撲は興味ありますか。
System もちろんですとも！千代の富士全盛期世代ですから。ちなみに好きな力士は誰ですか？
User 世代が同じかもしれませんね。私も千代の富士から貴乃花時代が好きでした。
System 確かに強い横綱がいましたね。個人的には北の湖とか若乃花なんかも好きでした。
User 北の湖は有名ですね。若乃花というと3代目のほうは良く知っています。
System おおー！初代貴ノ花じゃないですか！若い頃イケメンだったんですよー。
User 初代貴乃花は確かにイケメンでした。その子供達も立派でしたね。
System それこそまさに平成の大横綱一家ですね！でも残念ながら長男の花田勝治さんは若くして亡くなってしまったんですよ。
User そうですね。力士は短命な方が多いですよ。
System やっぱ厳しい世界だからでしょうかね。体脂肪率一桁台なんてザラらしいですもんね。
User ホント見た目じゃわからないですよ。
System いやほんとすごいですよね。僕ならすぐ病気になっちゃいますよ。おっと、もうそろそろお時間ですね。今日はありがとうございました！
User こちらこそ、楽しい時間をありがとうございました。
System またぜひよろしくお願ひします！

対話の分析

- 対話破綻につながる発話の分類（アノテータ1名が付与）
 - 前回コンペのNTTCS破綻分析 (杉山ら, 2020) から新たなカテゴリとして**スタイル変化・情報量過多**を追加
 - 頻出の誤りは**矛盾・話題の飛び・事実誤認・スタイル変化**（特にスタイル変化による破綻度合いが高い）
 - キーワード関係誤り・概念誤り・対象違いが0** → 大規模汎用言語モデルの強み

破綻度

- B1: 微かな違和感
- B2: 明らかな違和感
- B3: 継続困難な破綻

誤りタイプ	定義・例	B1	B2	B3	自然	追隨	提供
既知	対話内に含まれている内容の質問	4	5	0	4.5	4.5	4.0
矛盾	既知質問以外の矛盾	7	7	1	4.1	4.0	3.7
話題の飛び	話題や話が不連続に遷移	8	11	1	4.6	4.6	3.8
キーワード関係誤り	キーワードの関係が誤り (例: 佐賀は長崎ですか?)	0	0	0	-	-	-
概念誤り	キーワードではないが関係が誤り (例: 有田焼はどんなところ?)	0	0	0	-	-	-
事実誤認	事実と異なる発話 (例: 有川浩の容疑者Xの献身)	17	8	0	4.0	3.6	3.6
対象違い	システム自信の発話に対して、相手話者が発話したかのように話す	0	0	0	-	-	-
勝手	勝手に対話を終わらせようとする	2	1	2	2.5	3.0	2.5
スタイル変化	タメ口や攻撃的になり、発話スタイルが変化	5	11	4	3.8	3.2	3.4
情報量過多	情報量が多く、まくしたてるような発話	4	2	2	4.0	4.0	4.5
その他	上記に含まれないが不自然な発話	32	9	0	4.2	4.0	3.8
合計		79	54	10			

対話の分析

- 対話破綻につながる発話の分類 (アノテータ1名が付与)
 - 前回コンペのNTTCS破綻分析 (杉山ら, 2020) から新たなカテゴリとして**スタイル変化・情報量過多**を追加
 - 頻出の誤りは**矛盾・話題の飛び・事実誤認・スタイル変化** (特にスタイル変化による破綻度合いが高い)
 - キーワード関係誤り・概念誤り・対象違いが0** → 大規模汎用言語モデルの強み

破綻度

- B1: 微かな違和感
- B2: 明らかな違和感
- B3: 継続困難な破綻

誤りタイプ	定義・例	B1	B2	B3	自然	追隨	提供
既知	対話内に含まれている内容の質問	4	5	0	4.5	4.5	4.0
矛盾	既知質問以外の矛盾	7	7	1	4.1	4.0	3.7
話題の飛び	話題や話が不連続に遷移	8	11	1	4.6	4.6	3.8
キーワード関係誤り	キーワードの関係が誤り (例: 佐賀は長崎ですか?)	0	0	0	-	-	-
概念誤り	キーワードではないが関係が誤り (例: 有田焼はどんなところ?)	0	0	0	-	-	-
事実誤認	事実と異なる発話 (例: 有川浩の容疑者Xの献身)	17	8	0	4.0	3.6	3.6
対象違い	システム自信の発話に対して、相手話者が発話したかのように話す	0	0	0	-	-	-
勝手	勝手に対話を終わらせようとする	2	1	2	2.5	3.0	2.5
スタイル変化	タメ口や攻撃的になり、発話スタイルが変化	5	11	4	3.8	3.2	3.4
情報量過多	情報量が多く、まくしたてるような発話	4	2	2	4.0	4.0	4.5
その他	上記に含まれないが不自然な発話	32	9	0	4.2	4.0	3.8
合計		79	54	10			

対話の分析

- 対話破綻につながる発話の分類 (アノテータ1名が付与)
 - 前回コンペのNTTCS破綻分析 (杉山ら, 2020) から新たなカテゴリとして**スタイル変化・情報量過多**を追加
 - 頻出の誤りは**矛盾・話題の飛び・事実誤認・スタイル変化** (特にスタイル変化による破綻度合いが高い)
 - キーワード関係誤り・概念誤り・対象違いが0** → 大規模汎用言語モデルの強み

破綻度

- B1: 微かな違和感
- B2: 明らかな違和感
- B3: 継続困難な破綻

誤りタイプ	定義・例	B1	B2	B3	自然	追従	提供
既知	対話内に含まれている内容の質問	4	5	0	4.5	4.5	4.0
矛盾	既知質問以外の矛盾	7	7	1	4.1	4.0	3.7
話題の飛び	話題や話が不連続に遷移	8	11	1	4.6	4.6	3.8
キーワード関係誤り	キーワードの関係が誤り (例: 佐賀は長崎ですか?)	0	0	0	-	-	-
概念誤り	キーワードではないが関係が誤り (例: 有田焼はどんなところ?)	0	0	0	-	-	-
事実誤認	事実と異なる発話 (例: 有川浩の容疑者Xの献身)	17	8	0	4.0	3.6	3.6
対象違い	システム自信の発話に対して、相手話者が発話したかのように話す	0	0	0	-	-	-
勝手	勝手に対話を終わらせようとする	2	1	2	2.5	3.0	2.5
スタイル変化	タメ口や攻撃的になり、発話スタイルが変化	5	11	4	3.8	3.2	3.4
情報量過多	情報量が多く、まくしたてるような発話	4	2	2	4.0	4.0	4.5
その他	上記に含まれないが不自然な発話	32	9	0	4.2	4.0	3.8
合計		79	54	10			

まとめ

- ・ 大規模汎用言語モデル **HyperCLOVA** をベースとした対話システムを構築
- ・ 汎用言語モデルだけでは解決できない、対話のドメイン課題に対するアプローチの導入
- ・ 対話システムライブコンペティション4のオープントラック予選で1位を獲得
- ・ **流暢な文章生成は可能になったので、効果的な対話制御にフォーカスする必要がある**

今後の展望

- ・ **倫理的な応答生成**を行うためのフィルタリング強化
 - ・ より高範囲な単語リストを収集するだけでなく、文意理解による非倫理的な応答の除去
- ・ 大規模汎用言語モデルにおける **Hallucination** (虚偽情報の生成) の回避方法を模索
 - ・ 知識ベースやペルソナ情報を正確に応答に反映させるには？
- ・ **対話戦略**を考慮するモジュールの導入
 - ・ 対話状態や対話行為に応じたプロンプトの切り替え

Appendix

Few-Shot Learning - 大喜利の例

大喜利をします。

お題: 13日の金曜日に暴れ回るジェイソン。14日の土曜日は何をしてる？

回答: それなりに反省している。

お題: ナルシスト漁師の特徴を教えてください。

回答: キャッチ&キス&リリース。

お題: 昼休みに弁当箱をあけたら「うわ、嫁怒ってるわ…」どうなった？

回答: この大豆は遺伝子を組み替えています。

お題: 0円でできる超スーパー暇つぶしとは？

回答:



HyperCLOVA

まぶたを高速開閉する。

ペルソナ維持応答生成プロンプト - 抽出

ショット例

[最後の発言から属性情報を抽出します。]

[会話]

A: あ、ところであなたの誕生日っていつですか??

B: わたしは11月25日ですよー。

A: あと血液型はなんですか?

B: AB型です。

[Bの属性情報] **血液型はAB型。**

ここをペルソナとして利用

- **HyperCLOVA**を用い、最後の発言からペルソナを抽出する
- 抽出対象の話者を”B”として、ユーザー・よつばの**両方のペルソナ情報を生成**
- 対話セッション中は**ペルソナを保存**しておく
- 毎ターナー番最後にペルソナを抽出・保存

