

考察：「未夢と香澄の改造計画」

いよいよ未夢と香澄の大型改版にかかる時がきました。彷徨と香澄が動いて、未夢プロジェクトとしては一段落したのです。いよいよ、未夢プロジェクトの成果を実践に耐え得るシステムに仕立て上げる好機が到来したのです。予定しているのは、

- (1) 永続化
- (2) 曖昧性処理の整理と充実化
- (3) 固有表現への対応
- (4) 各種文法表現への対応(実践化)

です。未夢プロジェクトの最後の山場と思います。後は、辞書の充実とか、文法の手直しとか、小さい修正作業のみになると予想しています。今年いっぱいかけて1万語くらい対応できるレベルにしていくつもりです。

1. 曖昧性処理の充実

1.1 曖昧性

曖昧性のクラスは以下の通り。

(1) 分かち書き化

「はははははは、ははのははははははとわらう」

->「母は八八八、母の母は八八八と笑う」

(2) 修飾、被修飾関係

「私が長野で買ったバッグを落としてしまった」

->「長野で」は「買った」に掛かるのか、「落とした」に掛かるのか。

(3) 多義語

->猫を呼ぶ(call)、猫を「ぬくぬく」と呼ぶ(name)

(4) 多機能語

->助詞「と」、「の」、「も」

(5) 格の決定

「リンゴが食べたい」

->「リンゴ」は主格でなくて目的格

(6) 格の補填、指示代名詞の解決

「リンゴが食べたい」

->「私は」を補填しないといけない。

(7) ステージの決定

a. 「信子と明雄の乗った列車は長野駅についた。」

b. 「明雄は長野駅で降りた。」

c. 「信子は缶コーヒーを飲みながら手を振った。」

->の文で、信子は長野駅にいるのか列車の中にいるのか曖昧である。

(8) シーンの決定

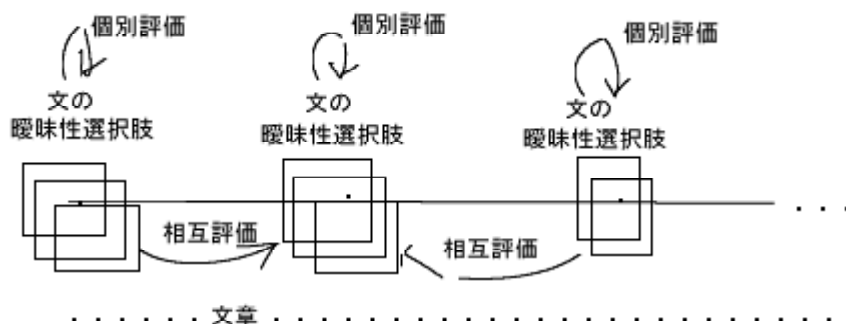
->ステージに曖昧性があるので、アクターのグループが曖昧になる。

1.2 曖昧性の解決

今まで、鍵と鍵穴という仕組みで汎用的に曖昧性を管理しようとしていましたが、どうもうまいかない。曖昧性クラスで異なった処理をしていくのが素直だと、一部コーディングしてきました。

でも、技術を修正して、鍵と鍵穴手法で解決していくことに戻すつもりです。鍵穴は曖昧性管理オブジェクトに曖昧カテゴリと曖昧の発生場所を指定して行う。カテゴリ別に曖昧処理メソッドを選択して、個別な処理をしていく。そのとき、重要なのは、

- (1) 意味記号の共起
 - (2) 意味記号ネットワークの構造情報とのマッチング
 - (3) 前後の文から得られる文脈記号と文脈記号の作るネットワークとの共起
- この3つを解析していくことです。



図：曖昧性管理、曖昧性低減評価構造

解析は弛緩法で順次曖昧性を取り除くという手法を取ります。また、文脈記号とそのネットワークはオンデマンドで意味解析をすることで得ていきます。オンデマンドでなくすべて一括解析でもいいでしょう。意味解析はコーパス、共起情報を基盤に行います。コーパスによって、単語では持たない、単語相互の意味関係を得ていく必要があるのです。例えば、「彼は彼女にチョコレートをあげた」という文では、「彼は彼女に好意を持っている」というような彼と彼女の意味関係を得ていく必要があります。共起情報、ネットワーク情報、文脈記号は専門の辞書をもって、システムの外部から解析処理を制御できるようにします。

(1) 分かち書き化

- ・ 単語の共起

stext	refer_target
ND(学校)	ND(生徒)

(2) 修飾、被修飾関係

・ 単語の共起

stext	refer_target
ND(学校)	ND(生徒)

・ コーパスマッチング

stext	corpus
VB(歩く)	@は*.ND(信子)@で*.ND(公園)@@

(3) 多義語

・ 単語の共起

stext	refer_target	mean
ND(呼ぶ)	XX(を)	call

・ コーパスマッチング

stext	corpus	mean
VB(呼ぶ)	@を*.ND(*)@と*.ND(*)@@	call

(4) 多機能語

・ コーパスマッチング

stext	corpus	mean
XX(と)	ND(*)_XX(と)_ND(*)	CF(と ; and)
XX(と)	VB(*)_XX(と)_VB(言う)	XX(と ; clause)

(5) 格の決定

- ・ 単語の意味記号の充実化（意味記号間の共起）とコーパス

stext	corpus	
VB(見る)	@が .agent.ND(*:+human) @は .agent.ND(*:+human) @を .direct_object.ND(*) @@	

- ・ 関係の共起

stext	corpus	mean
VB(与える)	agent->target,agent->direct_object	+favorite

(6) 格の補填、指示代名詞の解決

- ・ 格コーパス

stext	corpus	
VB(見る)	@が .agent.ND(*:+human) @は .agent.ND(*:+human) @を .direct_object.ND(*;+favorite) @@	

- ・ 関係の共起

stext	corpus	mean
VB(与える)	agent->target,agent->direct_object	+favorite

(7) ステージの決定

- ・ 単語の内包関係

stext	value	character	large_object	small_object
授業	*	{先生,生徒}	{教室,椅子,机,黒板}	{ノート,鉛筆,白墨}

2 . 永続化

RDB の一つのテーブルで曖昧性と文章シーケンスを管理していきます。ですから、レコードとして、曖昧候補と作業情報を分けて行かれるようになっていなければいけません。

id	s_no	a_no	text
separate	000001	000001	ND(信子 :+human) XX(は) ND(学校 :+location+object) XX (^)VB(行く :+move)
modify	000001	000001	(ND,0,1 信子 :+human) (XX,1,4,は) (ND,2,3,学 校 :+location+object) (XX,3,4,^) (VB,4,-1,行く :+move)
semantics	000001	000001	(ND,0,1 信子 :+human) (XX,1,4,は) (ND,2,3,学 校 :+location+object) (XX,3,4,^) (VB,4,-1,行く :+move); EpisodeSet(VB,4;(agent.*.0;location,to,2;modifier,*,5))
animation	000001	000001	Stage(time,*,*,*;location,ND,100,学 校) Scene(100,2) Cut ((ND,0,1 信子 :+human) (XX,1,4,は) (ND,2,3,学 校 :+location+object) (XX,3,4,^) (VB,4,-1,行く :+move); EpisodeSet(VB,4;(agent.*.0;location,to,2;modifier,*,5) ;subcutelement())

3 . 固有表現

次のような表現をサポートしていきます。

- (1) 未定義語の英字をシンボルとして扱う。意味記号は"+symbol"。
- (2) 未定義語の中に数値があれば数名詞として扱う。意味記号は"+number"。
- (3) xx 首相、xx 県、とかある程度人名、地名とわかる単語は名詞として扱う。また、意味記号も生成する。
- (4) 「リンゴを3つください」というような表現をサポートする。

【固有表現用辞書】

stext	mean
首相	+human
公園	+location
月	+time
個	+number

おわり