

# チャット対話における発言間の継続関係と応答関係の同定\*

徳永泰浩 乾健太郎 松本裕治  
奈良先端科学技術大学院大学 情報科学研究科  
{yasuh-to,inui,matsu}@is.aist-nara.ac.jp

## 1 はじめに

省略補完や代用表現の解釈といった対話理解のための対話構造のモデル化と解析は、音声対話を対象にした機械翻訳の分野で特に重要とされている。これに対し、チャット対話を対象とした対話構造のモデル化と解析は、情報抽出やコミュニケーション支援といったチャット対話を言語資源として利用する研究においても重要とされている [3, 4, 6]。

チャット対話では、表 1 のようにメッセージを送受信することで対話が進む。対話は文字データとして記録されるため、そのまま言語資源として利用できる。しかし、チャット対話に既存の対話構造モデル [7] をそのまま適用することは難しい。表 1 の 25 と 27 の発言<sup>1</sup>のように、質問とそれに対する応答のような意味的につながりを持つ発言が隣接しない場合がある。また、質問とそれに対する応答を構成する発言自体も 31 と 32, 33 の発言のように区切って送信 (区切り送信<sup>2</sup>) される場合がある [11]。

Khan ら [3] や小倉ら [6] は、チャット対話を対象に、意味的につながりを持つ発言の同定実験を行なった。彼らは、同定に必要な手がかりとその効果について調査しているが、対話構造の解析には触れていない。高梨ら [8] は独話を対象に節の同定を試みた。表 1 の例で言うと、31 と 32, 及び 32 と 33 の発言間にあるような節の同定を試みたが、発言 33 のような倒置の同定には触れていない。

我々は、チャット対話における発言間の関係を同定することを目的とし、チャット対話の特徴を考慮した対話構造のモデルを提示し、このモデルに基づいた対話構造を同定する手法を提案する。具体的には、表 1 の 31 と 32, 33 の発言のように、区切り送信によって複数の発言に分割された同一話者による発言間の二項関係を継続関係、継続関係にある発言のまとまり間にある質問と応答のような発言間の二項関係を応答関係と定義し、これらの関係に基づいて、発言をまとめあげることで対話構造を解析する。本研究では、この問題のある発言とそれと先行する発言との間に継続関係または応答関係があるか否かの 2 値分類問題に分解し、コーパススペースの教師あり機械学習を試みた。解析対象はオフラインのチャット対話ログである。

本稿では、本手法による対話構造の解析結果と、継続関係の同定における倒置の同定結果について報告する。

## 2 対話構造のモデル

我々は、Sinclair ら [7] の対話構造モデルを拡張し、チャット対話に適用した。表 1 の対話例に提案するモデルを適用した結果を図 1 に示す。

Sinclair らは対話構造を複数の階層から構成される対話のやりとりの構造としてモデル化している。まず、対話中の発言を対話行為と見なす。次に、同一話者による複数の対話行為を質問やそれに対する応答などの対話の

表 1: チャット対話例

ID	話者	発言
...	...	...
24	A	もう少し早くやるべきだったのでは
25	B	仕事ですか？
26	C	すみませぬ m( )m
27	A	いえ
28	A	ファンハール解任の話です
29	C	インテル戦前にですか？
30	A	いよいよ解任されそうって話です
31	B	ファンファールのサッカーを
32	B	出来る面子じゃありませんよ
33	B	バルセロナは
34	C	確かに
...	...	...

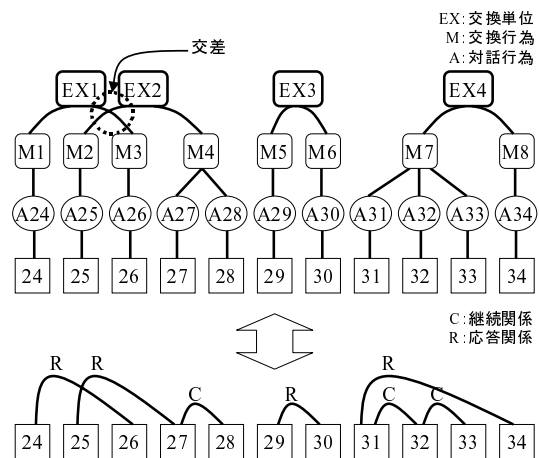


図 1: チャット対話の対話構造モデル

同所構造の中での機能を表現する交換行為にまとめる。この交換行為のまとまり (例えば、質問と応答のペア) を交換単位とし、異なる話者の発言の間を表現する。

しかし、チャット対話における発言を既存の対話行為と同様に見なすことは難しい。既存のモデルでは、発言を基本単位としており、区切り送信によって区切られた表 1 の 31 のような個々の発言を対話行為と見なしていない。我々は、このような発言を発言権の保持を意図する対話行為と見なし、モデルの基本単位を発言単位に変更する。また、発言 24 と 26 のように関連する発言が隣接しない現象を表現できるよう、交差を許すモデルに拡張する。

我々が提案するモデルでは、各発言を対話行為と見なし、継続関係にある対話行為のまとまりが交換行為に対応し、応答関係にある交換行為のまとまりが交換単位に対応する。つまり、対話構造を解析することは、発言間の二項関係を同定していく作業とみなせる。

## 3 実験用コーパス

本研究では、教師付き機械学習を利用し、チャット対話の対話構造解析を行なう。そのために、訓練データとして対話構造を付与したコーパスが必要になる。しかし、こ

\*Identification of Relevant Utterances in Chat Conversations  
TOKUNAGA Yasuhiro, INUI Kentaro and MATSUMOTO Yuji  
Nara Institute of Science and Technology

<sup>1</sup>チャットでの基本単位は発言単位よりも一段小さな単位となる。本研究ではこの単位を発言と定義し発言と区別する。

<sup>2</sup>区切り送信は発言権を相手に渡さないために起こるとされている。

れまでの研究では主に人手で作った規則を用いて解析する手法をとっていたため、訓練データとして使える大規模な対話構造付きコーパスは存在しない。そこで我々は、対話構造を付与したコーパスを作成した。

### 3.1 コーパスの元にしたデータ

本研究で使用するコーパスは、IRCで公開されているチャット対話データを独自に収集し、それを元に作成したものである。使用したチャットシステムは、CHOCOA<sup>3</sup>である。CHOCOAでは、分単位の受信時間、送信者名、送信内容の3つ組が一回に送受信できる。本研究では、これを入力対話の基本単位とし、付属の機能を用いて、対話データを収集した<sup>4</sup>。

収集した対話データは、2人対話 34 対話 (5180 発言) と 3人対話 35 対話 (6725 発言) の合計 69 対話である。発言者の異なりは 27 名であり、多くの話題は進路、サッカー、テレビゲームについてである。収録期間は約 4 カ月である。全発言に占める関係先を持つ発言数は表 2 の通りである。

表 2: 全発言に占める関係先を持つ発言数

	2人対話	3人対話
継続関係先をもつ発言数	1635	1599
応答関係先をもつ発言数	500	589
総発言数	5180	6725

### 3.2 対話構造の表現方法

対話構造を発言間の二項関係で表現する。表 3 に示すとおり、各発言の関係先の発言 ID を明示する。関係先の発言が 2 つ考えられる場合 (例えば、2 つの質問発言に対して 1 発言で答える) は両方の発言を関係先とする。本研究では、5 項目 (受信時間、送信者名、送信内容、発言 ID、関係先 ID) を訓練事例の基本単位とした。

表 3: 対話構造の表現例

ID	関係先	話者	発言
...	...	...	...
31		B	ファンファールのサッカーを
32	31	B	出来る面子じゃありませんよ
33	32	B	バルセロナは
34	31	C	確かに
...	...	...	...

コーパスの信頼性をタグ (関係先 ID) の一致率で評価した。評価尺度として Kappa 値 (Kappa Statistics) を採用した。コーパスから取り出した 400 発言について 2 人の作業者がそれぞれタグを付与し、Kappa 値を求めた結果は 0.892 であった。結果から、評価対象として信頼性、一貫性のあるデータであり、本研究の評価対象として有効であることが確認された。

## 4 提案手法

### 4.1 対話構造解析アルゴリズム

入力対話に対し、対話の先頭の発言から最末の発言にかけて順番に解析する。提案した対話構造の階層モデルに従い、まず継続関係の同定を入力対話全てに対して試み、発言を交換行為にまとめあげる。その後、交換行為間の応答関係を同定し、交換行為を交換単位にまとめあげる。

<sup>3</sup><http://www.labs.fujitsu.com/freesoft/chocoa/>

<sup>4</sup>対話参加者の入退出の通知などチャットシステム自体が通知する発言は分析対象から除外した。

#### 4.1.1 継続関係の同定

発言間の継続関係を同定することは、発言を交換行為にまとめあげることである。入力対話に対し、対話の先頭の発言から最末の発言までを走査しながら、その位置での発言 (以下、対象発言と呼ぶ) ごとに次のステップに従って処理が進む。

##### ステップ 1: 継続先発言候補の抽出

継続関係にある発言が対象発言に隣接しているとは限らないため、対象発言から  $n$  発言前までの対象発言と同一話者の発言とのペアを全て抽出する。一つの対象発言に対する  $n$  個の異なる先行発言とのペアの集合を候補ペア集合とする。

##### ステップ 2: 継続先発言の決定

分類器は、表 4 に示す素性集合を利用して、候補ペア集合に含まれるすべてのペアに対して継続関係があるか否かを分類する。複数のペアが関係ありと分類された場合は、分類器の出力するスコア (確信度) の一番高いペアに一意に決定する。候補のペア全てが関係なしと分類された場合、対象発言には継続関係にある先行発言が存在しないとする。つまり、対象発言は交換行為を構成する先頭の発言と同定される。

表 1 の対話を例に継続関係の同定手順を図 2 に示す。図は対象発言が 28、先行発言の参照数  $n$  が 4 の場合を表している。まず、対象発言と同一話者による発言である 24 と 27 の発言とのペアを候補ペアとして抽出し、分類する。分類後、発言 24 と 28 及び 27 と 28 のペアと複数のペアが関係ありと分類されるが、スコアを比較した結果、27 と 28 のペアが継続関係となる。

#### 4.1.2 応答関係の同定

交換行為間の応答関係を同定する。交換行為間の関係を同定することは交換行為を交換単位にまとめあげることである。入力対話に対し、対話の先頭の交換行為発言から最末の交換行為までを走査しながら、その位置での交換行為 (以下、対象交換行為と呼ぶ) ごとに次のステップに従って処理が進む。

##### ステップ 1: 応答先交換行為候補の抽出

応答関係にある交換行為が対象交換行為に隣接しているとは限らないため、対象交換行為から  $n$  交換行為前までの対象交換行為と異なる話者の交換行為とのペアをすべて抽出する。一つの対象交換行為に対する  $n$  個の異なる先行交換行為とのペアの集合を、候補ペア集合とする。

##### ステップ 2: 応答先交換行為の決定

分類器は、表 4 に示す素性集合を利用して、候補ペア集合に含まれるすべてのペアに対して応答関係があるか否かを分類する。複数のペアが関係ありと分類された場合は、分類器が出力するスコア (確信度) の一番高いペアに一意に決定する。候補のペア全てが関係なしと分類された場合、対象交換行為には応答関係にある先行交換行為が存在しないとする。つまり、対象交換行為が交換単位を構成する先頭の交換行為と同定される。

### 4.2 分類器

本手法では、素性として複数の発言の形態素情報の 2 つ組の組合せを全て考えるため、素性の異なりは数万となり分類器への入力ベクトルは高次元になる。そこで、分類器としてサポートベクタマシン [10] を利用した。スコ

表 4: 素性と抽出対象

継続	抽出対象	CRRu, PREu, NANu_s, NBNu_s
	素性	単体 発言 (CRRu, NBNu_s) の先頭の形態素, 発言 (NBNu_s 以外) の末尾の形態素, 発言の末尾の表層表現 二項間 発言間の結末度, CRRu と PREu 間の発言時間の差
応答	抽出対象	CRRm, PREm, NANm_s, NANm_d
	素性	単体 交換行為の対話行為クラス, 交換行為の末尾の表層表現 二項間 CRRm と PREm 間の名詞及び動詞の一致, CRRm と PREm 間の発言時間の差

CRRu: 対象発言, CRRm: 対象交換行為, PREu: 先行発言, PREm: 先行交換行為, NANu\_s: 対象発言と先行発言の間にある対象発言に最寄りの同一話者の発言, NANm\_s: 対象交換行為と先行交換行為の間にある対象交換行為に最寄り同一話者の交換行為, NANm\_d: 対象交換行為と先行交換行為の間にある最寄りの対象交換行為と異なる話者の交換行為, NBNu\_s: 対象発言の発言時間から 1 分以内に発言された最寄りの対象発言と同一話者の発言.

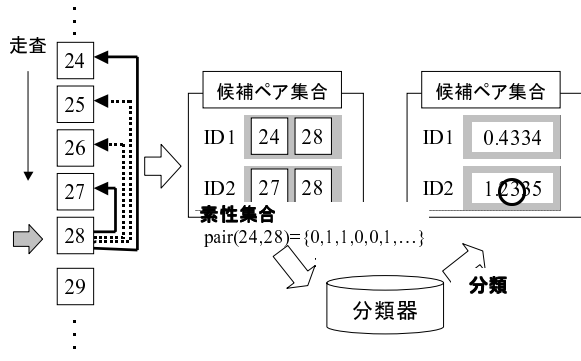


図 2: 継続先発言の決定例

ア(確信度)としては分離平面からの距離を用いた. 訓練時には, 継続関係及び応答関係の同定処理のステップ 1 と同じ処理を行ない獲得した各ペアから表 4 に示す素性集合を抽出し, 各訓練事例とする.

### 4.3 素性

発言の末尾の表層表現:

素性の抽出対象となる各発言の末尾の表層表現が句点または読点, クエスチョンマークであるか否かの 2 値を素性とする.

CRRu と PREu 間の発言時間の差:

CRRu と PREu 間の発言時間の差が 2 分以上離れているか否かの 2 値を素性とする.

発言間の結末度<sup>5</sup>:

CRRu と PREu 間及び CRRu と NBNu\_s 間の結末度を計り, どちらのペアの結末度が強いのかの 2 値を素性とする. 結末度の強さは, 個々のペアの  $\langle n(\text{名詞}), rel(\text{助詞}), v(\text{動詞}) \rangle$  の共起確率値  $P(\langle n, rel, v \rangle)$  の大きさで計る.  $P(\langle n, rel, v \rangle)$  を推定する手法として単語の共起を潜在的な意味からの同時発生とみなす Probabilistic Latent Semantic Indexing(PLSI)[9] を使用し, 共起確率モデルを作成した. 発言間の共起確率モデルは任意の入力  $\langle n, rel, v \rangle$  に対して, 共起確率値を出力する. モデルの訓練事例には新聞記事を用いた<sup>6</sup>.

交換行為の対話行為クラス:

素性の抽出対象となる各交換行為が各交換行為クラス(以下, クラス)を持つか否かの 2 値を素性とする. クラスは, あらかじめ定義した表層表現パターンとクラスの対応を示すテーブルを用いて推定し, クラスラベルを用いて表現する. クラスは, 発話行為タグ [1] を参考に 20 種類を用意した. このテーブルを対話行為辞書<sup>7</sup>と呼び, 内容の一部を表 5 に示す. 複数のクラ

表 5: 対話行為辞書(一部)

表層表現パターン	クラスラベル
はい	[肯定・許諾][あいづち]
ほう	[あいづち]
できるの?	[真偽情報要求]

スラベルが記述されているものは, 行為の多義があることを示す.

交換行為の末尾の表層表現:

素性の抽出対象となる各交換行為を構成する発言の中で一番最後の発言の末尾の表層表現がクエスチョンマークであるか否かの 2 値を素性とする.

交換行為間の発言時間の差:

CRRm のと PREm の先頭の発言の発言時間の差が 5 分以上離れているか否かの 2 値を素性とする.

## 5 評価実験

### 5.1 実験条件

実験には, 本研究で作成したチャット対話コーパスを用いた. 事前にすべての発言に対して茶筌 [5] を利用し形態素解析を行なった. 結果の妥当性・再現性を高めるために 5 分割交差検定を行なった. 素性の組合せも同定規則として採り入れるため, 分類器(サポートベクタマシン)のカーネル関数として 2 次の多項式カーネルを用いた. 参照する先行発言数は解析時及び訓練時共に, 継続関係で 7, 応答関係で 8 発言とした. これはコーパスにある発言から最も遠い関係先までの発言数である<sup>8</sup>.

### 5.2 調査項目

本手法及びベースラインの手法による対話構造の解析精度について調査した. ベースラインの手法では, 入力対話の先頭から最末の発言までを走査しながら, (ステップ 1) 対象発言と対象発言に最寄りの同一話者による先行発言が継続関係にあるとみなし, (ステップ 2) 正しい継続関係先を与えた状態で, 対象交換行為とそれに最寄りの異なる話者の先行交換行為が応答関係にあるとみなし, 対話構造を解析する.

継続関係と応答関係の同定実験の結果は, 再現率及び精度, F 値で評価した. 対話構造全体の解析結果は一致率で評価した. 関係先が複数ある発言では, いづれかひとつと一致した場合を正解とした.

$$\text{再現率} = \frac{\text{分類器が関係同定に成功した事例数}}{\text{実際に関係を持つ事例数}}$$

$$\text{精度} = \frac{\text{分類器が関係同定に成功した事例数}}{\text{分類器が関係を持つと判断した事例数}}$$

$$\text{一致率} = \frac{\text{本手法で判断した関係先がコーパスと一致した事例数}}{\text{コーパス中の発言の総数}}$$

<sup>5</sup> 発言同士の交換行為単位へのまとめやすさ

<sup>6</sup> 新聞記事 19 年分(毎日新聞 9 年分, 日経新聞 10 年分のべ 25,061,504 文)から 3 つ組  $\langle n, rel, v \rangle$  を抽出した.

<sup>7</sup> チャット対話コーパスから約 800 発言を抽出し作成した.

<sup>8</sup> 解析時に対象発言からいくつ先の発言までとのペアを考えるかという参照先数のパラメタ最適化の問題があるが, 今回は扱わなかった.

### 5.3 実験結果

本手法とベースラインの手法で解析した対話構造を比較することにより、本手法の有効性を評価した。

対話構造全体の解析結果を表 6 に示す。BL2p 及び BL3p は 2 人対話及び 3 人対話データにベースラインの手法を適用した結果、提案 2p 及び提案 3p は 2 人対話及び 3 人対話データに本手法を適用した結果である。一致率は、2 人対話及び 3 人対話それぞれにおいて、本手法がベースラインの手法を上回っている。

表 6: 対話構造の一致率

	BL2p	提案 2p	BL3p	提案 3p
一致率	.334	.859	.265	.847

継続関係及び応答関係の分類器の評価結果を表 7, 表 8 に示す。継続関係にある発言の多くは隣接しているため、BL2p, BL3p の再現率は高い。しかし、互いに隣接する発言でも実際には継続関係を持たない場合もあるため、精度は低い。これに対し本手法では、継続関係を持たない発言間の関係も同定しているため、精度も高くなっている。応答関係の同定精度は、BL2p, BL3p に比べ本手法が大きく上回っている。交換行為に対話行為クラスを付与することにより、質問への適切な応答を同定できたと言える。以上の結果から、本手法の有効性が確認できた。

表 7: 継続関係の同定結果

	BL2p	提案 2p	BL3p	提案 3p
精度	.381	.844	.348	.843
再現率	.985	.919	.986	.891
F 値	.549	.880	.515	.867

表 8: 応答関係の同定結果

	BL2p	提案 2p	BL3p	提案 3p
精度	.098	.740	.131	.765
再現率	.613	.466	.682	.531
F 値	.169	.572	.220	.627

素性の組合せを変えて実験を行なったところ、(i) 発言 (NBNU<sub>s</sub>) の先頭の形態素及び (ii) 発言間の結束度の素性が倒置された発言との継続関係を同定する際に効いていることを確認した。

素性の効果を検証した結果を表 9, 表 10 に示す。提案 2p は全ての素性を採り入れた結果、2p-12 は (i)(ii) の素性両方を除いた結果、2p-1 は (i) の素性を取り除いた結果、2p-2 は (ii) の素性を取り除いた結果である。(i)(ii) の素性を両方とも採り入れた結果は、精度、再現率、F 値共に他の条件の結果よりも高い。

素性 (i) が倒置の同定に効くと考えられる理由を事例 (1) を用いて説明する。(c) の発言の先頭の形態素は接続詞であり、新しい交換行為の始まりを表している。「現在の自分の交換行為が終わる前に、新しい交換行為を始めない」と仮定すると、発言 (b) は発言 (a) に対する倒置である可能性が高いと判断できる<sup>9</sup>。

- (1) a) A: 一年だとすぐって感じしないですよ  
 b) A: その先生のツテで探せるし  
 c) A: また、大学の空を探せませすよね?

素性 (ii) が倒置の同定に効くと考えられる理由を事例 (2) を用いて説明する。個々のペアの共起確率値  $P(\langle n(\langle$

<sup>9</sup>実際には、言いさし(発言を途中で中止すること)の可能性もあるが、コーパス中に数件の事例しかなかったため、今回は扱わなかった。

表 9: 素性の組合せと継続関係の同定結果

	提案 2p	2p-12	2p-1	2p-2
精度	.844	.821	.805	.807
再現率	.919	.868	.875	.915
F 値	.880	.844	.838	.858

表 10: 素性の組合せと継続関係の分類毎の再現率

	提案 2p	2p-12	2p-1	2p-2	発言数
助詞類	.942	.912	.911	.936	1021
接続詞類	.879	.846	.851	.869	215
特殊表記	.666	.666	.666	.666	6
名詞類	.919	.887	.887	.919	62
倒置	.754	.500	.568	.676	102
挿入	.698	.714	.693	.693	49

名詞), *rel*(助詞), *v*(動詞)) を計ると、〈宛, 来〉の方が〈宛, 直す〉よりも大きい。このため、発言 (b) は発言 (a) に対する倒置である可能性が高いと判断できる。

- (2) a) A: メール来ちゃった  
 b) A: 個人宛で  
 c) A: 前の文章を直す

## 6 おわりに

本研究では、チャット対話を対象として対話構造解析を行なった。本研究の成果は以下の通りである。

チャット対話の対話構造モデルの提案: チャット対話の対話構造モデルを提示し、このモデルに基づいた対話構造の同定手法を提案した。このモデルは、継続関係と応答関係という発言間の二項関係で対話構造を表現する。対話構造を付与したコーパスの作成: 対話構造を付与したコーパスを作成した。2 人の作業者同士のタグ付けの一致度を調査した結果は、Kappa 値で 0.892 であった。結果から、本コーパスは評価対象として信頼性、一貫性のあるデータであることが確認できた。

チャット対話の対話構造解析: 対話構造全体の一致率は 2 人対話 0.859, 3 人対話 0.847 であり、ベースラインの手法を上回っている。また、先行研究 [8] で難しいとされた倒置の同定を試み、倒置の同定に効果のある素性を特定した。

## 参考文献

- [1] 荒木 雅弘, 伊藤 敏彦, 熊谷 智子. 発話行為タグ標準化案の作成. 人工知能学会誌, Vol.14, No.2, pp. 53-62, 1999.
- [2] 細馬 宏通. チャットは何を前提としているか -チャットの時間構造と音声会話の時間構造-. 「身体性とコンピュータ」: bit 別冊, 共立出版, 2000.
- [3] Faisal M. Khan, Todd A. Fisher, Lori Shuler, Tianhao Wu, William M. Pottenger. Mining Chat-room Conversations for Social and Semantic Interactions. 2002.
- [4] 倉林 則之, 山崎 達也, 湯淺 太一, 蓮池 和夫. ネットワークコミュニティにおける関心の類似性に基いた知識共有の促進. 情報処理学会論文誌, vol.43, No.12, 2002.
- [5] 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 高岡 一馬, 浅原 正幸. 日本語形態素解析システム『茶室』version2.2.9 使用説明書. 2002.
- [6] 小倉 加奈代, 石崎 雅人. チャット対話における関連発言同定のための表層情報の分析. 人工知能学会 SIG-SLUD-A203-P05, 2003.
- [7] Sinclair, J. McH, Coulthard, R.M. Towards an analysis of discourse. In *Advances in spoken discourse analysis*, Routledge, 1992.
- [8] 高梨 克也, 丸山 岳彦, 内元 清貴, 井佐原 均. 話し言葉の文境界-CSJ コーパスにおける文境界の定義と半自動認定-. 言語処理学会 第 9 回年次大会, 2003.
- [9] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 50-57, 1999.
- [10] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
- [11] Werry, C.C. Linguistic and interactional features of Internet Relay Chat. In S. C. Herring(Ed.) *Computer-mediated communication: Linguistic, social and cross-cultural perspectives*, pp. 47-63, 1996.