情 報 学 基 礎　66－8
デジタル・ドキュメント　32－8
（２００２．３．１５）

第１回国際ワークショップ"NLP and XML"の概要とマルチモーダル・デジタル・ドキュメントのISO標準について

野村直之*1　中挾知延子*2　Key-Sun, Choi *3

*1 法政大学　*2 東洋大学
*3 KAIST (Korean Advanced Institute of Science and Technology)

概要：　豊かな構造と意味・論理表現を備えた XML 文書が広範に広まるための前提として、自然言語で書かれた文書の(半)自動化や、半構造文書の構造化、自動要約・自動編集の必要性が認識されている。また、自然言語処理の高度化のためには、何れの研究手法、方法論をとるにせよ、大規模なタグ付き言語コーパスや、その利用技術の開発、標準化が重要、との共通認識がある。さらに、自然言語生成技術で近年主流となりつつある３段パイプライン方式等の rapid prototyping で複数の研究者が XML の有用性を確認しつつある。

これらの共通認識を広め、技術開発、標準化を進展させる目的で、先般、第１回国際ワークショップ"NLP and XML"が開催された。http://hal2001.itakura.toyo.ac.jp/~chiekon/nlpxml/　本稿では、"NLP and XML"にて発表された大学、企業の関連技術を総覧するとともに、特に重要な国際標準化動向として、術語定義マークアップ言語標準 ISO/TC37/SC3 の現状と展望について報告する。

An Abstract of the 1st International Workshop on NLP and XML with a Special Emphasis on ISO/TC37/SC3 Standard of Multimodal Document

Naoyuki NOMURA*1　Chieko NAKABASAMI*2　Key-Sun Choi*3
*1 Hosei Univ.　*2 Toyo Univ.　*3 KAIST (Korean Advanced Institute of Science and Technology)

Abstract: XML, the universal structured data representation meta-language, has become the standard framework for publishing on the net, as well as the standard e-commerce language to build B2B and B2C Web services. A major concern for this scenario is the "point of creation" bottleneck, at which creating useful, well-structured XML data can consume unduly amount of time and effort. Hopefully, NLP should be able to resolve this bottleneck by automating the conversion from unstructured or semi-structured text data into XML documents with much richer structure hidden in the original NL descriptions. This is "NLP for XML" that can give some intelligence, or disambiguation capabilities to XML generating engines. Conversely, XML can help NLP researches, especially the ones with annotated corpus based approaches, by providing them with the knowledge representation frameworks for morphological, syntactic, semantics and/or pragmatics information structure of NL resources. In many cases, XML should be able to provide NLP with deeper semantic structure clues and thus realize much more robust, higher precision NLP applications.

The vision described above has led to the 1st International Workshop on "NLP and XML," which is summarized in this paper.　ISO/TC37/SC3 standard for terminology mark up is briefly mentioned as well.

1 はじめに

XML が次世代 Web、構造化情報処理の決定版となることが確定しつつある中、データベースの研究コミュニティ、情報検索の研究コミュニティが XML を基盤とした技術体系の再構築に大きくシフトしつつある。自然言語処理の研究コミュニティでは、1990 年代初頭以来、大規模テキストコーパスの構築と活用、事例主導の解析・生成と規則主導の解析・生成との融合を課題としてきた。そして、構文解析結果が通常木構造によって表現されることから、XML や XML データベースを研究ツールとして活用する潜在需要が大きかったといえる。これが"XML for NLP" である([Nomura&Nakabasami2001])。この立場では、単にツール、実装の手段として XML を用いるだけでなく、データやアルゴリズムの見通し良いモデルを構築するのに XML による表現を積極的に活用しようという流れが最近出てきている。特に言語生成や文章要約のアーキテクチャが目につく[Seki2001]。このアプローチは、木構造を扱う高品質で網羅的な関数を備えた XML の各種処理系が誰でも無料で使えることから、設計と実装を表裏一体と出来る点で優位性をもつと考えられる。

一方、他の分野、産業界からの要求として、「レガシー文書を XML に[半]自動変換」というものがある。これに象徴されるのが"NLP for XML" である。あらゆるネットワーク上のデータが相互運用可能な XML 文書に統一され、快適な利用が出来るというバラ色のストーリーは結構だが、データの意味構造をタグ付けする辛い作業を 100%人間がやるのか？ 語彙が少なく、あまり厳密なシンタックスが要求されない HTML でさえ苦労しているのに。という問いに対する１つの強力な回答が、自然言語処理によって有用且つ均質、高品質な XML 文書をプレーンテキストや HTML 文書の蓄積から[半]自動生成する、というものである。

以上の問題意識に賛同するメンバーが集まり、第１回国際ワークショップ"NLP and XML"は、NLPRS2001 環太平洋自然言語処理国際シンポジウム本大会終了の翌日 2001 年 11 月 30 日に、北欧、西欧、東欧、北米、アジア各国から 40 名以上の参加者により盛況の内に開催された。９本の一般講演の間に"Aspects of Semantic Annotation"と題して橋田浩一氏が招待講演を行った。これは、「当面(今世紀中？)は自然言語の完全な意味理解は困難であるから、機械に対して予め実在の意味構造を与え、その都度必要なだけ記述した構造を漸進的に保持していく。さらにその構造をマルチモーダル化することで実社会にとって有用なアプリケーションを提供しつつ研究を進めよう。」という趣旨のものであった。自ら発案し主導する GDA (Global Document Annotation; http://xml.coverpages.org/ni2001-09-18-a.html) のみならず、自然言語によるインデックスが注目される MPEG7 等の重要性がアピールされた。

以下、一般講演の概要を逐次紹介しつつ、当日の質疑を踏まえた著者によるコメントを記す。筆者の１人 Key-Sun Choi が座長を務める術語マークアップの国際標準 ISO/TC37/SC3 については、講演６ "A Common XML-based Framework for Syntactic Annotation" [Ide2001] の中で取り上げる。

2 NLP & XML ワークショップ一般講演の概要

"Pipelines, Templates and Transformations: XML for Natural Language Generation"
自然言語生成のための XML – パイプライン、テンプレート、変換
Graham Wilcock (Univ. of Helsinki)

本講演は、自然言語生成に用いられるパイプラインアーキテクチャ、テンプレートならびに XML による中間表現間の変換の実装を XSL、DOM、Translet (XSL を Java クラスにコンパイルしたもの) で行い、それぞれの性能を比較したものである。性能比較の結果として、パイプライン実装に要求される XML 間の変換においては、Translet が最も適切であると述べられている。理由として、(1)XSL は手軽に実装ができるが、会話の応答が返される度に同じスタイルシートを再読み込みしなければならないため時間がかかるので実用的でない、(2)DOM では生成される木の各ノードの属性、タイプ、値というような細かなプログラミングが必要で、Java のプログラミングスキルが要求されるため、限られた者だけにしか用いられないため、と述べられている。

この XML ベースの自然言語生成技術は、ヘルシンキ大学の USIX-Interact と称する大きなシステム実証研究の一貫として開発中のものである。図1に USIX-Interact のシステム構成を示す。今回の発表の位置付けは、図1の右側部分 Presentation Manager における言語生成部の制御に関するものである。USIX-Interact は領域知識、タスク知識に基づいて自然な音声対話の実現を目指すものである。Input Manager, Dialogue Manager に含まれる多種の処理エンジンとの間で柔軟かつ高機能、高性能な連携を実現するのに XML が貢献するものと期待される。

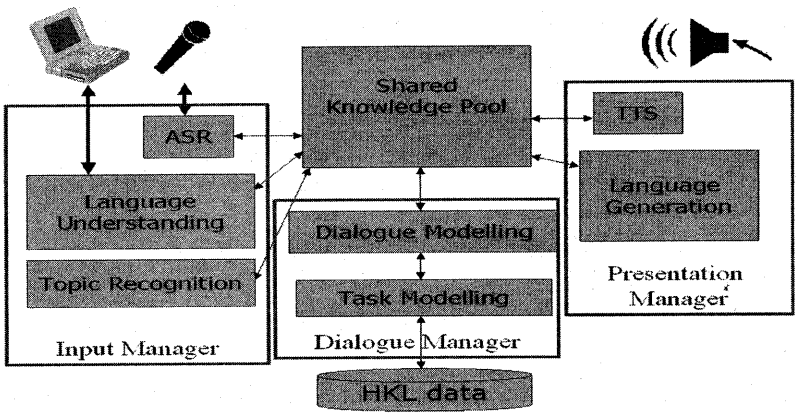

図1　ヘルシンキ大学 USIX-Interact のシステムアーキテクチャ

"Dialogue Scenario Generation from XML-based database"
XML データベースからの対話シナリオの生成
Masahiro Araki, Kiyoshi Ueda, Takuya Nishimoto, and Yasuhisa Niimi (Kyoto Inst. of Technology)

本講演では、対話処理に用いるための VoiceXML の半自動生成システム VOX が提案されている。VOX システムでは、ユーザからの検索要求に対して XML 形式で貯えられた対話データベースの中から一致するものを抽出し、それらを VoiceXML に変換する。ユーザは音声を通じて VoiceXML に変換されたデータを受け取ることができる。システムの利点として、情報発信側ではユーザの要求に対して想定される多様な応答例を作る必要がなく、細かな応答パタンの違いを XML 形式にすることで吸収させてしまうことができることである。VoiceXML という音声メディアの活用形態を提案した研究である。
VoiceXML は、比較的初期の XML 応用言語であり、他のシナリオのソースに制御を跳ばす<goto>タグをもつなど、何でも強引に書けてしまう側面をもつ。このため、直接人手でコーディングするのでなく、本論文のように、知識ベースと何階層かの高次のモデルを元に、一定の規範に則して自然な対話を「生成」する、というのは適切なアプローチであるといえよう。

"Natural Language Enabled Web Applications"
自然言語を柔軟に扱うための Web アプリケーション
Kuansan Wang (Microsoft)

本講演では、Web アーキテクチャと XML をベースにしたマルチモーダルなプラニングシステムが提案されている。これからの Web アプリケーションでは、ユーザに対する要求と応答については柔軟な自然言語と多様なデバイスへの対応をせざるを得なくならないであろう。ユーザからの要求については、音声認識のメカニズムを備えた自然言語処理技術が必要であり、データの表現形式として XML での言語データへの意味アノテーションは適切なフレームワークを提供する。一方で、XML はコンテンツとプレゼンテーションの分離を実現でき、XSLT などを使って多様な出力メディアに対応できる。

口頭では割愛されたが、論文中ではプラニングシステムといったケーススタディを示しながら、次世代のWebアプリケーションに自然言語処理技術とXMLの適用が不可欠であると述べられている。図2に示す(音声)言語解析によるユーザ主導 WebService 統合実験は、WebService を自然言語によるリクエスト（音声入力）を文脈解析し、Semantic Schema (XML)を検索・照合し、リアルタイムで提供されるサービスをオン・ザ・フライで結合して提供しようというものである。

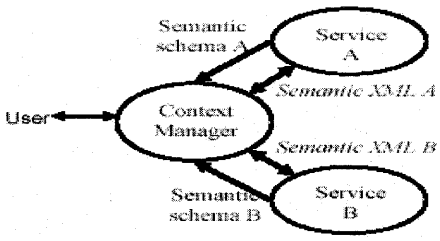

Figure 2: Federated Understanding. User's utterance is parsed into partial semantic XML and sent to related Web services based on their semantic schemas. Service A and B do not have to be designed to work with each other.

図2　Microsoft .NET 構想における自然言語対話による WebService 自動結合

複数のアクションの指示を内包する自然言語文から WebService を検索し接続性、セマンティクスを検証し、実際に結合する。その結果、例えば、「その日の為替レートで決済するカード会社とつながっているカメラの価格比較&販売サイトで日本円で送料込み３万円以下のデジカメが欲しい」という問いに対して必要十分な回答を返せることを追求する。しかし、早期の実用化を指向して、ヒトの言語理解を模したような高度な自然言語解析能力に偏重せず、浅い意味解析、文脈解析にとどめている。代わりに、テキストのレイアウトやレトリック構造、オントロジー構造に相当するスキーマ定義や、ひな形に沿ったレイアウト生成モデルによる複数スキーマの結合、そして、そのためのルール記述言語を重用している。つまり、実際によく使われる意味表現の構造モデルや、意味表現からの連携サービス生成モデルを強力にすることによって、自然言語解析の負担を大幅に軽減しようという発想で、自然言語入力による頑健な WebService 結合を実現しようとしている。

なお、前講演と同様、本論文も VoiceXML を痛烈に批判している。曰く、データと制御、書誌情報が渾然一体となり、しかも、ＸＭＬで知識を記述する良さ＝「宣言的（静的）知識で書きやすい」を捨て去っている、と。代案として、（1）名前空間等 XML 本来の美点を活かした最小限の記述による機能拡張を指向、（2）WSDL への代案たる Semantic Schema 自体に拡張性をもたせる、という階層化されたモデルによるアプローチを提案している。解決案の面からも、論文２と同様の立場をとっているといえよう。

"XML and NLP: Their interaction and their role for HLT Applications"
XML と NLP の相互作用とその役割　(HLT = Human Language Technology)
Declerck Thierry & Peter Wittenburg (DFKI & Max Planck Institute)

本論文では、筆者が所属する MUMIS (Multimedia Indexing and Searching Environment) プロジェクトにおけるマルチメディアコンテンツでの情報検索システムの紹介を通じて、XML と NLP が相互に果たす役割が述べられている。MUMIS はサッカーのビデオにおいて XML でのアノテーションを付けることにより、たとえば「ゴールシーンの中で成功したゴール」のビデオの 1 場面を抽出することなどを容易に実現できるようなアプリケーションの開発を目指している。論文中ではそのようなマルチメディアアノテーションのベースとして MPEG7 が有力であると言及されており、Automatic Hyperlinking と称する XML と NLP

の相互作用について論じられている。つまり、XML は NLP アプリケーションにおいて各フェーズにおける処理を容易にし、フェーズ間の自然言語データのパイプ役を果たす一方で、NLP は XML でのアノテーション技術に言語的な手がかりと制約を与えることができるとしている。そしてこれらの相互作用はナレッジマネジメントに重要な役割を担うことが述べられている。その１つのポイントは、図３に示されるように、領域知識（イベント構造）を表す木構造と言語知識を表す木構造を XML により柔軟にマッピングしよう、という点にあるようだ。

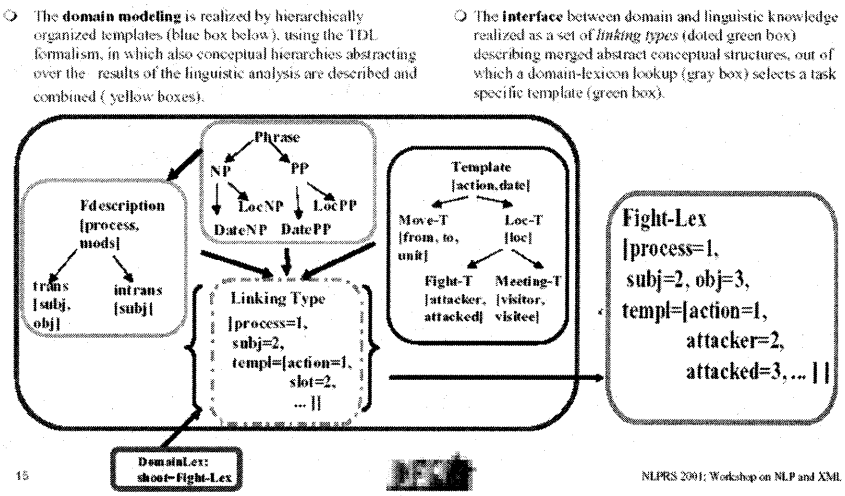

図３　領域知識と言語知識を結びつけるパラダイム（DFKI）

"Discussion Mining: Knowledge Discovery from Online Discussion Records"
ディスカッションマイニング — オンラインのディスカッション記録からの知識発見
Akiko Murakami, Katashi Nagao, and Koichi Takeda (IBM)

本論文では、Web の BBS に投稿された記事について、Discussion グラフを作って XML で内容を記述することにより、話題ごとのサマリを生成する手法が提案されている。またグラフの構造を解析することにより、投稿者の中での Authority Person と Hub Person を同定する方法も示されている。論文では投稿された記事の引用部分に着目し、それを手がかりに記事間を関係付け Discussion グラフを生成する。生成されたグラフはビジュアルインタフェースを備えており、各記事間の関係を把握しつつ、記事のメッセージに GDA を用いてタグを付けていくことができる。タグ付けされたメッセージはマイニングのためのインプットとなり、話題ごとのサマリを生成するために用いられる。もう 1 つの Authority Person の同定にはメッセージの引用率などの定量的なスコアが用いられている。

本論文は、XML により、形式と内容の両方を手がかりとして同時に活用したマイニング技術を紹介したものである。メール文書のヘッダー部分の構造と本文テキストの解析結果の構造とを GDA で融合した結果（図４）、有為な知識発見が見通しよく出来るようになった、と評価されよう。

```
<?xml version="1.0" encoding="Shift_JIS" ?>
 <gda id="Katashi_Nagao4">
  <SUBJECT>Re: ホイール(20V)お返事</SUBJECT>
    <DATE>2001/11/21 16:14:18</DATE>
    <AUTHOR>Katashi_Nagao</AUTHOR>
    <BODY>
      <QUOTE ref="572-5 572-5 572-6">-
            <adp>
          -  <np 読み="でぃーら―" 品詞="普通名詞">ディーラー</np>
          -  <ad 読み="に" 品詞="格助詞">に</ad>
          -  </adp>-
```

図4 GDAに準拠してメール文書を構造化した例

"A Common XML-based Framework for Syntactic Annotation"
統語アノテーションのためのXMLを用いたフレームワーク
Nancy Ide, Laurent Romary, and Tomaz Erjavec (resp. Vassar College, Campus Scientifique, & Jozef Stefan Institute)

本論文では、筆者らの開発している XCES アノテーションフレームワークについての概要と統語アノテーションへの適用が述べられている。XCES は EAGLES (Expert Advisory Group on Language Engineering) ガイドラインの一部で、タグ付き言語コーパスにおけるデータ構造の標準フォーマットを提案するものである。XCES はすでに Terminological Markup Framework (ISO16642) などにも採用されている。本論文のフレームワークを用いることで、アノテーションを行う際にそれぞれの目的に合った内容に具体化できることが意図されている。論文ではフレームワークの適用例として、異なった形式の統語アノテーション間の比較ができることと、統語パーザの評価が行えることが示されている。XCES は特定の統語アノテーションに依存しないスケルトン構造、データカテゴリレジストリなどのリソースと、特定のアノテーションに依存する部分からなる。XCES において重要なコンポーネントとして Annotation Markup Language (AML) が提案されており、異なる統語アノテーション間の比較を行うための中心的役割を果たすものである。

本論文でのフレームワークはタグつきデータの生成や、それを活用する汎用ツールの開発も可能にすることが示唆されている。Brants らが開発中の統語アノテーションの一例 Negra の例を示し(図5)、ISO16642 の有用性が主張された。しかし、その検証、ならびにメタモデルの洗練は今後の課題といえる。

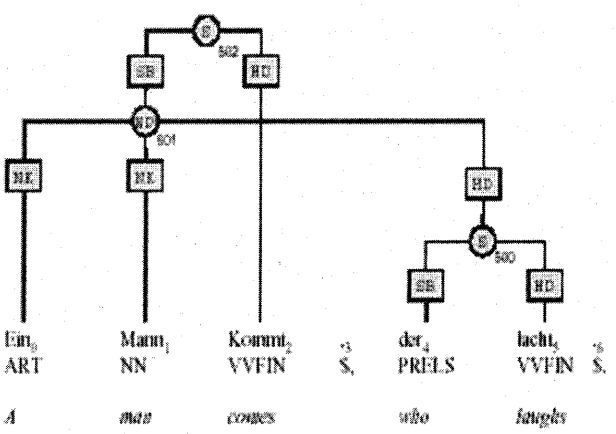

図5 ISO16642準拠のNegraによる統語アノテーション

"XML-Based Linguistic Annotation of Corpus"
XML タグ付き言語コーパスの実践 – 医学・生物学言語コーパスのアノテーション事例
Jin-Dong KIM, Tomoko OHTA, Yuka Tateishi, Hideki Mima, and Jun'ichi Tsujii

本論文では、医学・生物学における論文アブストラクトの言語データへのアノテーション事例が報告されている。アノテーションは人手で行われており、医学・生物学分野の言語リソースを構築する目的としたGENIA プロジェクトの一環として行われている。 GENIA の言語リソースは、オントロジー、専門用語、タグ付きコーパスから成り、その中でも重要な位置を占めるものがタグ付きコーパスである。コーパスの作成のために、プロジェクトでは GENIA Project Markup Language (GPML) を開発し、GPML の中でのアノテーションのための XML タグセットは XLiNo と呼ばれている。現時点では GENIA の言語リソースは MEDLINE データベースから抽出した 670 の論文アブストラクトについてアノテーションが行われており、論文では XLiNo の概要とアノテーションの具体例(図６)が示されている。

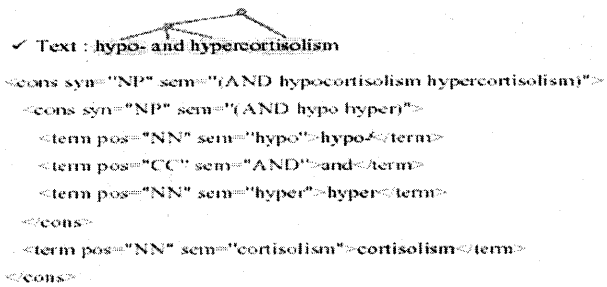

図 6 等位接続を表す XLiNo (XML tag set for Linguistic anNotation)

"A Proposal on Information Hiding Methods using XML"
XML を用いた情報隠蔽手法の提案
Shingo Inoue, Kyoko Makino, Ichiro Murase, Osamu Takizawa, Tsutomu Matsumoto, and Hiroshi Nakagawa (resp. Mitsubishi Research Institute, Communications Research Laboratory, Yokohama National Univ. and Univ. of Tokyo)

本論文では、XML ドキュメントの特徴を利用して暗号を埋め込む情報隠蔽の手法が提案されている。手法としては、XML 文書の文法構造を利用したものと、XML 文書の論理構造を利用したものに分けられている。いずれも暗号を埋め込むことで XML 文書の意味を変えてしまうものではない。XML の表示には XSL などがよく使われ、コンテンツとプレゼンテーションが一致している必要がないので、暗号を埋め込んで送信するためには適していると述べられている。XML 文書の利用方法としての新たな可能性を示唆する研究である。「Canonical XML で空白類など些末的な違いが同一視されてしまったらどうするのか?」という質問への回答は、「その場合は、当該箇所を避けて秘密情報を埋め込む」というものであった。

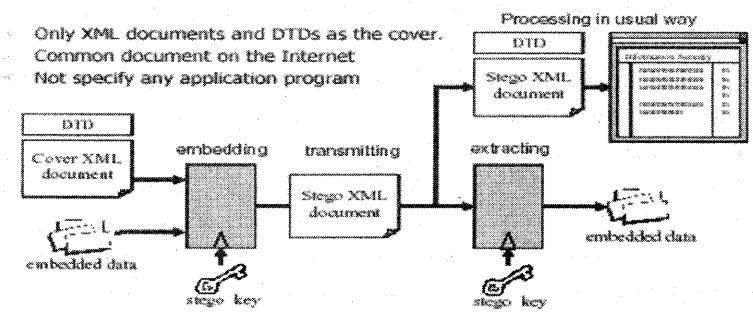

図7　XML 情報隠蔽のモデル

"A Preliminary Study of Lexical Density for the Development of XML-based Discourse Structure Tagger"
XMLによる談話タグ開発のためのレキシカル密度の研究
Lawrence Y. L. Cheung, Tom B. Y. Lai, Benjamin K.Tsou, Francis C.Y. Chik, Robert W. P. Luk, and Oi Yee Kwong (City Univ. of Hong Kong & Hong Kong Polytechnic Univ.)

本論文では、裁判の判決における会話について XML 化し、タグの情報を談話構造におけるカテゴリの推定に適用したものである。論文ではあらかじめ“冒頭陳述”、“事実の説明”、“論拠”、“判決”という４つのカテゴリを設け、会話のどの場所が、一つのカテゴリから別のカテゴリに移るところであるのかを同定する。そのために法律用語の辞書と、さらにその中から判決によく用いられる用語を集めた辞書が用意されている。語の出現頻度に基づきレキシカル密度を、LD_{GLT} $(s_i)=(N_{GLT,si} / N_{si}) * \log_{10}(N_{si})$ と定義する。これは１文中に法律語彙が占める割合を GLT(General Lexical Terms)に比して正規化したものである。レキシカル密度の変化率の大きな箇所が文書中のセグメント境界、すなわちカテゴリが変化するポイントに近いもの、との仮説を立て（図８）、それが検証された、とする報告である。レキシカル密度は、会話は段落番号、文番号や法律用語であるかどうかを示すタグによって XML 化されており、談話分析に XML のタグを利用している。

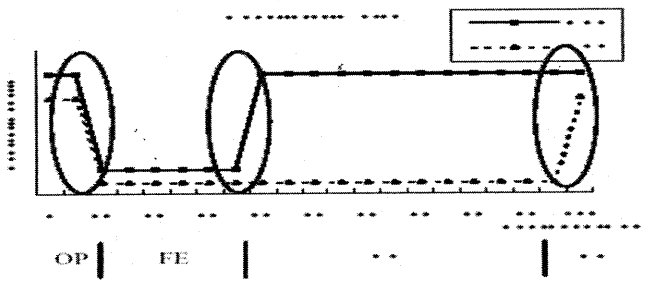

図８　文書中のセグメント境界をレキシカル密度の変化率から推定する手法

3　おわりに

以上、第１回国際ワークショップ"NLP and XML"の概要を紹介した。組織委員会、プログラム委員会の当初の予想に比べ、"XML for NLP"に大きく片寄りはしたが、ISO/TC37/SC3 の活動をはじめ、様々な構造記述の標準の提案や、相互運用性の高いアーキテクチャの提案が一同に介し、有意義な議論になったと考える。今後、XML により頑健且つ柔軟になった自然言語処理技術によって(半) 自動で付加価値を付けながら既存文書資産を XML に変換すること、そして、高精度の翻訳、要約等の実用システムの開発が期待される。

参考文献（下記で断り無いものは全て Proceedings of 1st Workshop on "NLP and XML," 2001）

[Nomura&Nakabasami2001] (the Workshop Proposal) http://hal2001.itakura.toyo.ac.jp/~chiekon/nlpxml/

[Seki2001] Seki, Y.and Harada K.,: "XML Transformation-based three-stage pipelined Natural Language Generation System" in Proceedings of NLPRS2001, Demonstration Session

[Hasida2001] Hasida, K.: "Aspects of Semantic Annotation," Invited Talk

[Wilcock2001] Wilcock, G.: "Pipelines, Templates and Transformations: XML for Natural Language Generation"

[Araki2001] Araki, M. et al.: "Dialogue Scenario Generation from XML-based database"

[Wang2001] Wang K.: "Natural Language Enabled Web Applications"

[Thierry2001] Thierry, D. et al.: "XML and NLP: Their interaction and their role for HLT Applications"

[Murakami2001] Murakami A. et al: "Discussion Mining: Knowledge Discovery from Online Discussion Records"

[Ide2001] Ide, N. et al.: "A Common XML-based Framework for Syntactic Annotation"

[Kim2001] Kim, D et al.: "XML-Based Linguistic Annotation of Corpus"

[Inoue2001] Inoue, S. et al.: "A Proposal on Information Hiding Methods using XML"

[Cheung2001] Cheung, L. et al. "A Preliminary Study of Lexical Density for the Development of XML-based Discourse Structure Tagger"