AFRL-RI-RS-TR-2009-262
Final Technical Report
November 2009

REASONING EFFICIENTLY FROM SELF-ORGANIZATION OF UNSTRUCTURED DATA (RESOUND)

HNC Software LLC

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th ABW, Wright-Patterson AFB Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2009-262 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/s/
JAMES M. NAGY
Work Unit Manager

/s/
MICHAEL J. WESSING, Deputy Chief
Information & Intelligence Exploitation Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
NOVEMBER 2009	Final	August 2006 - September 2009

4. TITLE AND SUBTITLE

REASONING EFFICIENTLY FROM SELF-ORGANIZATION OF UNSTRUCTURED DATA (RESOUND)

5a. CONTRACT NUMBER
FA8750-06-C-0200

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
31011G

6. AUTHOR(S)

Richard Rohwer

5d. PROJECT NUMBER
CASE

5e. TASK NUMBER
00

5f. WORK UNIT NUMBER
04

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

HNC Software LLC
3661 Valley Centre Drive
San Diego, CA 92130-3317

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/RIED
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
N/A

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TR-2009-262

12. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PA# 88ABW-2009-4790

13. SUPPLEMENTARY NOTES

14. ABSTRACT

During the two years since its effective start date (28 Aug 2006), the HNC IARPA CASE project has brought us closer to the goal of a universal and optimal approach to information extraction. Building on the earlier IARPA NIMD project, new algorithms were developed for unsupervised learning of hierarchical feature sets for text and imagery, and the Text Analysis Engine (TAE) SOA component of the CASE Integrated Architecture was extended to several languages and more thoroughly hardened and tested. Most importantly, we have clarified our understanding of universal abstract principles that can guide future research on information extraction and organization directly to its greatest potential payoff.

15. SUBJECT TERMS

Text Analysis Engine, CASE, Distributed Alignment, Association Grounded Semantics, Hierarchical Learning, Semantically-Driven Segmentation, Text extraction, Image identification, unsupervised learning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			James M. Nagy
U	U	U	UU	30	19b. TELEPHONE NUMBER (*Include area code*) N/A

TABLE OF CONTENTS

LIST OF FIGURES

1 INTRODUCTION

This is the final report of the research and development carried out under AFRL contract FA8750-06-C-0200, awarded to HNC Software LLC, then a unit of Fair Isaac Corporation (now FICO), effective 28 August 2006, and subcontracted to SRI International after their acquisition of most of the HNC business unit on 13 January 2009. This contract was funded by the IARPA CASE program, which was terminated after 2 of an originally planned 4.5 years.

The objective of this project is to build out certain key components of a vision for machine cognition technology capable of autonomously organizing and reasoning about data of any modality, including text, images, and video. This technology is to be integrated into the CASE system to thoroughly capture tacit knowledge in a usable form for data integration, collaboration support, and geospatial reasoning, and to provide dramatically improved capabilities for hypothesis generation and tracking.

The research was aimed at developing very general algorithms for the extraction and organization of information from unstructured sources such as natural language text and imagery, and creating deployable implementations of these algorithms. The desired functionality included automated unsupervised discovery of informative feature hierarchies in text and image data; i.e., discovery of words, phrases, and gramatical constructions in text and discovery of edges, objects, etc., in imagery, based only on statistical information obtained from unsupervised analysis of the data of interest. It also included alignment of terminology across languages and data sources, as well as alignment of visual features across different lighting conditions, scales, and orientations. These capabilities are fundamental requirements for developing tools to support intelligence analysis that can automatically and quickly adapt to the many diverse data sources involved. Enhanced named entity disambiguation capabilities were also desired.

Despite the shortened time frame, these goals were substantially met. In particular, substantial progress was made in learning visual feature hierarchies, cross-language terminology alignment, and named entity disambiguation.

2 BACKGROUND

The research carried out on this contract was largely an outgrowth of earlier work by HNC funded under the IARPA (then DTO) NIMD program. Fundamental principles were discovered during that work that clarified how to accomplish unsupervised information extraction and organization in general. The first of these was Association-Grounded Semantics (AGS), a mathematically sound formulation of a long-standing principle for capturing the meaning of a data object from the statistics of its usage contexts. To this was added Semantically-Driven Segmentation (SDS), whereby meaning carrying units such as single word elements and multi-word elements could be discovered from raw byte streams without applying any prior language knowledge. This principle largely guided the work on imagery feature learning in the CASE project. Finally a principle called Distributional Alignment (DA) was discovered which made it possible to align terminology between languages or dialects without using any language knowledge.

From these foundations, algorithms were developed for semantic clustering of terms and documents, unsupervised part of speech tagging, named entity recognition (NER), named entity disambiguation, relationship extraction, and several other natural language processing tasks. Several of these algorithms were packaged into a Services Oriented Architecture (SOA) called the Text Analysis Engine (TAE). These principles and technologies formed the starting point for the CASE work. Further information can be obtained from the Principle Investigator Meeting technical papers delivered to IARPA.

2.1 Technical Approach

The technical approach is framed within a scientific theory that is uniquely capable of resolving the fundamental conundrums of machine intelligence with which the CASE research areas were inextricably entangled. We begin the Technical Summary by highlighting the key concepts of this theory. With this background material in place, we then narrow the presentation to the main technical foci of this project:

- terminology alignment
- ontology alignment
- information extraction from video
- Their applications to data integration, collaboration, and geospatial reasoning

The science characterizes "meaningful structure" in a media-agnostic way, thereby supporting a bytes-to-thoughts vision of self-organizing cognitive technology, heavy on (mostly off-line) computation but light on demands for manually created data.

The scientific theory is built on three key concepts: Association-Grounded Semantics (AGS) by which meaning is captured mathematically via autonomous statistical processing, Semantically-Driven Segmentation (SDS) whereby semantically rich structure is automatically discovered, and Distributional Alignment (DA) for discovering abstract commonalities between different domains of discourse such as different topics, languages, or ontologies.

This project was centered on developing and exploiting the SDS and DA principles to automatically align terminology to improve collaboration effectiveness, interpret obscure languages and dialects, and align ontologies for database and knowledgebase integration or interoperation. It contains options that expand the effort into the media of images and video, demonstrating the universality of the underlying principles while developing new video and cross-media information extraction and interpretation capabilities. It also offers options to organize information for insertion into knowledge-bases, and to perform semantically-enhanced reasoning over knowledge-bases.

Our technical approach is built around the far-reaching but common-sense insights that understanding operates on ***meaning***, and that meaning derives from usage. This insight also emerges from a careful mathematical formulation of the concept of "meaning" in terms of Bayesian states of knowledge and Information Theory. This confluence is both re-assuring and enabling. It definitively tells us how to obtain a computationally usable grasp on what another institution means by its jargon, or how to organize massive data with respect to its semantics. It is the only way to keep up with continually evolving dialects such as those of "blog-ese", where (for now) "gr8" has come to mean "great" and "411" has come to mean "information". Similarly, in certain subculture dialects, "wedding party" is a deliberately obfuscated expression for "bombing".

Our method for deriving meaning from usage captures ***tacit knowledge*** in a form that can be incorporated into the lowest-level "subliminal" operations of a cognitive system, so that higher-level operations are automatically informed by it. Representing tacit knowledge implicitly at this level is a far more robust and unbiased approach than representing it explicitly at a higher level using standard Knowledge Engineering methods. Besides, explicit codification of tacit knowledge is wildly impractical. The problem is that tacit knowledge is trivial, and trivial information is copiously abundant. The number of facts one can state decreases markedly with their profundity, and conversely, increases limitlessly with their triviality. Tacit knowledge is the most trivial kind, so trivial to human experience that it is normally applied unconsciously. Therefore any attempt to catalogue all tacit knowledge in a high-level declarative form is guaranteed to fail. Yet many are reluctant to abandon such quests because they reach a Sisyphean illusion of imminent completion; whatever the latest question is, the knowledge-base is missing "just a few trivial facts" needed to answer it.

2.1.1 The mathematical theory of meaning

In research stretching back many years, but more recently intensified under the DTO NIMD program, we have formalized this basic insight as "**Association-Grounded Semantics (AGS)**". AGS enables us to derive meaning representations automatically from unbiased, language-agnostic statistical analysis of *unlabeled* corpora. Figure 1 illustrates the basic concept of AGS, the use of automatically-derived probability distributions to represent meaning. For any meaning-carrying unit (a ***lexeme***) such as a term in natural language or a feature in an image, one performs a statistical analysis over a corpus representative of the lexeme's usage to obtain a corresponding probability distribution. The distribution is defined over properties of the contexts in which the lexeme is mentioned. This set of contextual properties, called the *grounding variable*, is chosen in part for technical convenience and in part to achieve desired semantic nuance. Variables as simple as "word to the left" are fairly effective. In this way, an ensemble of lexemes (a ***lexicon***) becomes an ensemble of probability distributions, each of which constitutes a point in the mathematical space of all distributions over the grounding variable. This space of distributions is called an ***information geometry*** because it has a rich geometric structure that encompasses and extends Information Theory in both its Shannon and Fisher variants. This information geometry supplies us with a mathematically precise concept of a ***semantic space*** in which semantic relationships become geometric relationships. The synonymy relationship, in particular, corresponds to geometric distance, so clustering lexemes in this space yields groups of similar-meaning terms. Figure 1 shows semantic clusters from declassified military cable data and Chinese data.

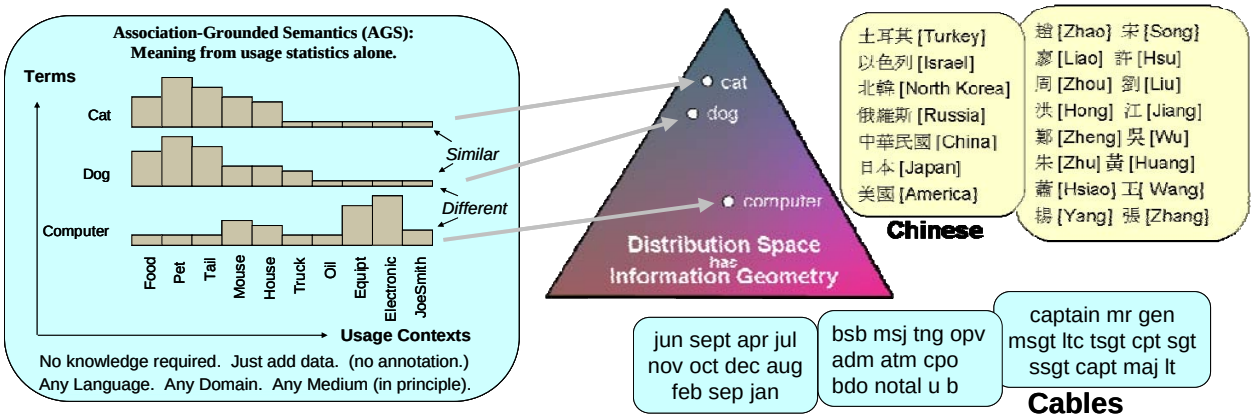
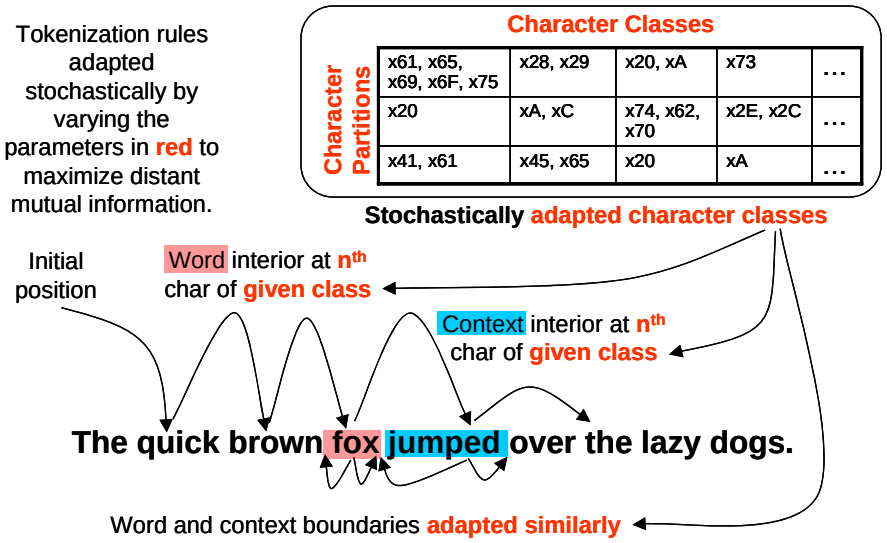


Figure 1: Association-Grounded Semantics (AGS)

In Figure 1, Data objects such as terms in a lexicon are associated with probability distributions describing the contexts of their mentions. These distributions are obtained by automated statistical analysis of corpora; no annotation is needed. These distributions form points in a high-dimensional mathematical space with rich geometric structure, information geometry, that mirrors the semantic content of the data objects (e.g., terms). So for example, clustering relative to the information geometry produces groups of semantically similar terms. This applies whether the data is well-formed English text, highly non-standard English such as military cable data, or text in a foreign language such as Chinese.

To apply AGS, one must choose a lexicon and a grounding variable. One might define white-space-separated strings to be lexemes, for example, or instead use stems formed by removing morphological affixes. Although common sense intuition has served as an adequately reliable guide in most cases, it would be better to have a rigorous criterion for optimizing the rules for "segmenting" lexemes from text (or images or other media). The information-theoretic framework of AGS provides the basis for formulating the desired principle, which we call Semantically-Driven Segmentation (SDS): The best lexicon encodes the most semantically rich information, where "semantically-rich" means "non-locally predictive". It is standard practice in image processing to optimize feature sets for maximal information capacity, but SDS holds that this criterion is deficient because it does not distinguish semantically rich structure from semantically poor noise, and that it is the distance over which features have predictive power that distinguishes the two. In exploratory experiments under NIMD, we established that applying SDS to raw byte-level data (that happens to be text) results in the spontaneous creation of tokenization rules for white-space-delimited terms, multi-word person names and morphological affixes. The principle has not been tested in other media. Figure 2 illustrates spontaneous derivation of a text tokenizer from application of SDS to a raw byte stream. A broad family of tokenization rules was defined based on scanning for bytes in a particular equivalence class to discover the interiors and boundaries of terms and contextual features. The equivalence classes and parameters of the scanning rules were varied in a stochastic search process aimed at maximizing an objective function expressing the SDS concept. Typical results of the tokenizer are shown on the right. We note a preponderance of informative multi-word units, often segmented at white space boundaries as humans would choose to do, and usually producing informative tokens.



No.	{variant} No. tokens ...
83	{Chief Executive} 35
82	{March 27 (PNS)} 10
	{March 30 (PNS)} 8
	{March 28 (PNS)} 8 ...
48	{Pervez Musharra} 20
46	{Prime Minister} 22
41	{Executive General} 26
	{Executive, General} 2
36	{General Perve} 16
36	{March 31 (PNS)} 8
	{March 16 (PNS)} 5 ...
30	{n the country} 17
	{nto the country} 1
26	{electoral rolls} 20
26	{Hurriyet Conference} 18
26	{Muslim League} 19
25	{enazir Bh} 22
24	{ammu and Kashmir} 7
22	{ the country} 8
21	{second phase} 18
20	{Munir A Sheikh} 15

Figure 2: Spontaneous derivation of a text tokenizer from application of SDS

AGS enables meaning comparison between lexemes that are mapped to distributions over the same grounding variable. This requires there to be a reasonably large set of contexts within which either lexeme may be mentioned. This is an onerous requirement in situations such as comparing words from different languages, because technically convenient context properties such as “word to the left” take values in separate languages, resulting in nearly or entirely disjoint sets of grounding values. One might fix the problem by applying a translation dictionary to the grounding values, but such hand-made resources introduce bias, expense and delay into the process of coping with new languages and evolving dialects. This is a poor idea, but it contains the germ of a most excellent one. Given any translation of the grounding values, the semantic space of one language can be mapped into that of the other, at which point the tools of information geometry can be applied to evaluate the translation. This being the case, the best possible translation can be derived automatically by solving a maximization problem. This is the principle we call Distributional Alignment (DA), illustrated in Figure 2, which includes a sample of term clusters aligned using our DA algorithm transLign. One can also think of DA as a process of rotating and reflecting one semantic space (in its typically hundreds or thousands of dimensions) so that a lexicon within it lines up as well as possible with a lexicon in another semantic space, as Figure 3 attempts to illustrate.

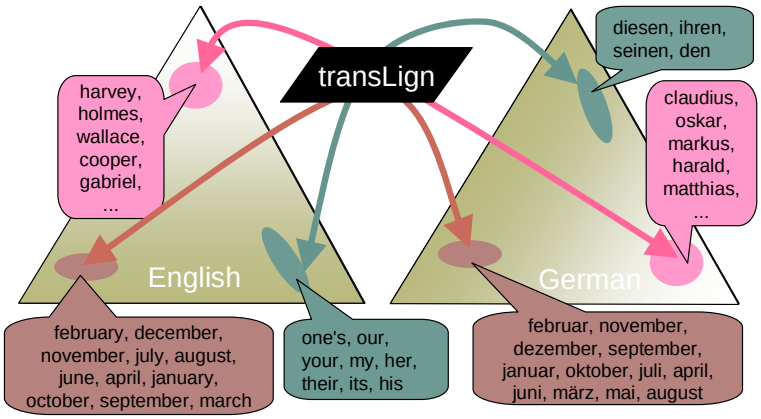


Figure 3: Distributional Alignment (DA)

In Figure 3, two different information geometries, i.e., spaces of distributions over different variables can be aligned by rotating and reflecting the (high-dimensional) spaces to match up the pattern of lexeme locations within them as closely as possible. This enables optimal semantic matching even if a shared set of contexts is not available.

Because DA aligns geometric shapes in a semantic space, it captures meaning at a level of abstraction independent of any particular language or dialect, and does so without recourse to tie-words or aligned corpora. It can also be used to find the best possible alignment between lexica that concern different subject matter, whether or not they differ in language. It achieves a level of abstraction unconcerned with the semantics of individual lexemes, but focused instead on the pattern of semantic relationships between the lexemes. This is the essence of analogy and metaphor. That such an important abstraction emerges naturally from the mathematics speaks volumes for AGS theory as the sound foundation of semantics, while also providing a practical approach to capturing this abstraction computationally.

Our mathematical theory of semantics, set in semantic spaces derived automatically using AGS to represent meaning, SDS to find the most meaningful structures, and DA to unify disparate semantic spaces abstractly provides a universal framework for data understanding, applicable in principle to any modality of unstructured data including text, images, video, and audio, or to structured data such as relational databases or knowledge-bases. Incorporation of these principles into Knowledge Engineering will impart semantic awareness to reasoning engines, enabling them to finally overcome the brittle, obsessively pedantic behavior that has always plagued this technology. Figure 4 summarizes this vision: A unified science and technology of semantics, knowledge, and reasoning, self-organizing massive data to achieve cognizant computing.

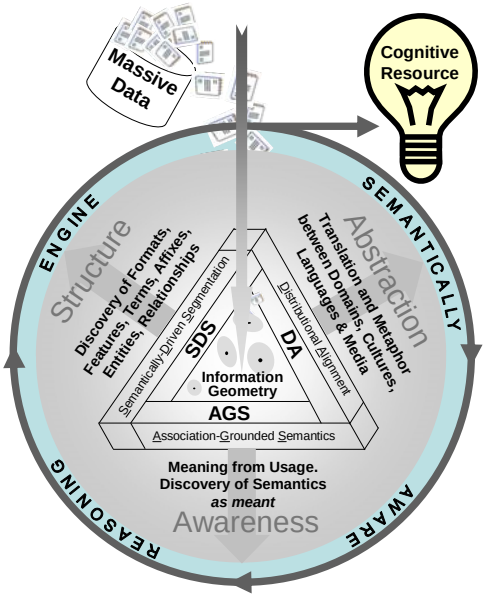


Figure 4: The scientific vision for deriving intelligence from statistical analysis of data

3 RESULTS AND ACCOMPLISHMENTS

Our research centered on **information extraction and organization**. Abstractly, information extraction can be viewed as feature set selection; meaningful features such as words or morphemes occur at particular locations in a sample of text, for example, whereas features such as objects, object parts, edges and corners appear at particular locations in images. Information extraction is the problem of learning not only a lexicon of feature values—the various words, object shapes, etc., but also the rules for detecting which feature values are present where. Information organization can be viewed as *hierarchical* feature set selection; collections of feature values are seen to have a "higher level" meaning when present in particular relationships. To accomplish information extraction and organization through unsupervised learning requires universal principles for assessing not only the amount of information expressed by any given (possibly hierarchical) feature set, but also the degree to which this information is *meaningful*.

We are following and fleshing out two such **principles**, both arising out of the earlier NIMD work, that we call Semantically Driven Segmentation (SDS) and Distributional Factorization or Alignment (DA). Both principles are developed from the broad principle that a *meaning* is a *pattern of usage*, a principle we call Association-Grounded Semantics (AGS). These principles are intimately connected with hierarchical feature building, because a pattern of usage of a feature value can always be represented as a probability distribution describing the ways in which it tends to be incorporated into higher-level features. It follows that the meaningfulness of a feature value can only depend on this distribution and properties of the higher-level feature syntax; this is the SDS principle. The question of which distributional and syntactic properties contribute to meaning, and how and why they do, remain central research issues, but the basic intuition is that a feature value is more or less meaningful to the extent that it participates in more or fewer higher-level patterns that have greater or lesser spatial extent.

Distributional Factorization and Alignment (DA) is a principle for separating independent factors of meaning, and a significant accomplishment of the CASE work was to show that this notion of "independence" is literally the statistical one. We have shown that at least for one striking example, the separation of language from word-sense meaning, DA can separate "semantic invariants" (uninteresting information) from "semantic content" (interesting information).

3.1 Image Feature Hierarchy Learning

We developed methods for unsupervised learning of a visual feature hierarchy. The lowest-level features are learned by vector quantization of (normalized) image patches at multiple scales, using a modified k-means algorithm to discourage learning of codebook vectors that are merely shifted translations of each other. Maximization over this feature set produces a level-1 symbol assignment. To form the second level, a co-occurrence consistency calculation is carried out to define a large set (typically about 14K) of syntactic combinations of level-1 symbols defining level-2 symbols. The co-occurrence consistency is determined by computing co-occurrence counts for all pairs of symbols in all possible syntactic arrangements within the grid, computing the pointwise mutual information for all of these, and growing combinations that all have high pointwise mutual information between themselves. These support a bag-of symbols description of images or image regions that can be used for image classification or retrieval. Level 3-features can be formed for relatively large regions of an image by vector quantization of these histograms. This process is illustrated in Figure 5.

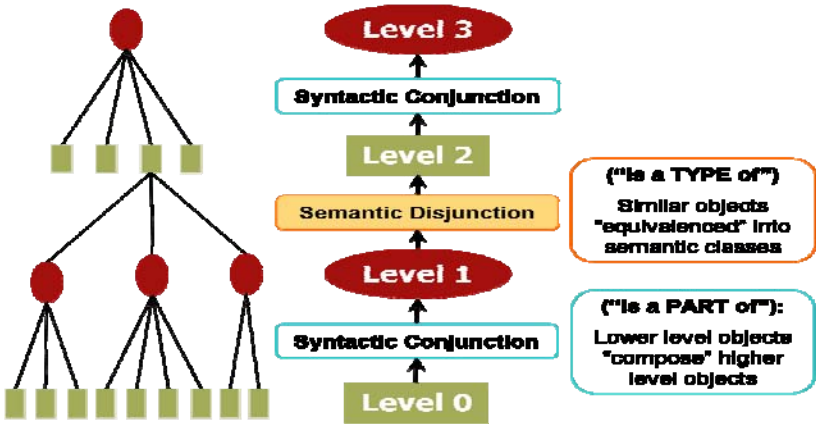


Figure 5: Hierarchical Learning Paradigm

We implemented baseline feature sets including a simple one based on Laws texture features (moments of 3x3 features in 32x32 patches) and a feature hierarch based on the S1, C1, S2, and C2 layers of visual cortex devised by Poggio's group at MIT[1]. We ran tests on the PASCAL Visual Object Classes Challenge 2007 (VOC2007)[2] data set (10K images in 20 classes) and Caltech256 data set (30K images in 256 classes). Using the features from these systems with a linear SVM classifier, our system gives somewhat better pair-wise classification results (by 4.8% Average Precision) than the Poggio system baseline on the VOC2007 data, which is typical of the unsupervised learning systems reported on that data set. The supervised systems do somewhat better.

[1] T. Serr, L. Wolf. S. Bileschi, M. Reisenhuber and T. Poggio, "Robust Object Recognition with Cortex-Like Mechanisms, IEEE Trans PAMI, 29, (2007).

[2] M. Everingham, L. Van Gool, C.K.I. Williams, J Winn, and A. Zisserman, PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

This feature hierarchy was defined over a uniform grid of image patches. Our intention was to further refine and deepen this hierarchy, and then move on to adaptive learning of the "interest" points defining the locations at which the features are defined. As discussed above, this co-adaptation of interest points (in text, these would be "mentions" or "tokens") and feature sets is the central issue in SDS. It is a difficult issue conceptually as well as computationally, because unfettered interest point adaptation would give total freedom for a feature learner to choose its data in trivial or otherwise counterproductive ways. It is therefore necessary to constrain this co-adaptation in a way that has little impact on the method's general applicability, but suffices to discriminate between "interesting" and "uninteresting" information in some useful sense. When the project was curtailed, we decided to conduct at least an exploratory experiment in this area.

We devised an objective function called "information contrast" for simultaneous adaptation of interest points and feature sets. We implemented the concept for imagery and obtained sensible exploratory results. We defined two-slot "predicates" in terms of any given placement of interest points in the images. Each predicate is instantiated once for each interest point, with the feature value at the interest point in its first slot. The second slot is filled by the feature value at the nearest neighboring point for one predicate, and the second-nearest neighboring point for the other. The objective is the difference between the mutual information for the second-nearest and nearest neighbor predicates. This means that interest points must be placed to make second-nearest neighbors highly predictable while making nearest neighbors unpredictable. (To reward making the nearest point predictable would make the problem too easy.) The idea is that second nearest neighbors will tend to be within the same object as the interest point, and nearest neighbors will tend to be on different objects. Visual inspection of results seems to bear this out.

We have not yet checked, but it would seem that this objective is also resistant to over-crowding of interest points into highly informative (but redundantly informative) regions, because a new interest point inserted into such a region would "steal" contrast from points already there, unless this region contained enough information contrast to support more interest points. Objectives of this nature have the merit of not being tied to any particular orthographic scale or informatic thresholds. Much work remains to be done, however to perfect these concepts and to integrate them with hierarchical feature learning.

3.2 Distributional Factorization and Alignment (Analogy detection)

We developed an improvement to the `transLign` algorithm developed in the NIMD program for unsupervised alignment of the lexica of different languages. The new algorithm, Distributional Factorization, directly produces mixed-language clusters rather than an alignment of pre-learned clusters in each language. This is a much more flexible procedure that overcomes the constraint of one-to-one alignment in particular. We created a highly optimized implementation and demonstrated good alignments between most European languages, using the EuroParl JRC corpus of over 20,000 documents translated into 22 European languages Cross-language information retrieval experiments were done to make quantitative comparisons for different parameter settings, language pairs, etc. Each document was matched to its translation by comparison of their distributions over 200 of mixed-language word clusters that were learned without using any cross-language resources or aligned corpora. German documents were retrieved from English at an F1 rate of 0.823, French from English at F1=0.881, and Hungarian at F1=0.046. The state of the art in this task had been F1=0.0, in the sense that such a zero-knowledge approach had not been considered feasible previously. It remains to study what fraction of the performance of more fully informed methods can be obtained from a zero-knowledge approach, and what applications in poorly resourced languages and idiomatic writing are enabled by this technology. We published the results in "Coarse Lexical Translation with no use of Prior Language Knowledge" at the Workshop on Building and Using Comparable Corpora at the Language Resources and Evaluation Conference (LREC) in Marrakech, Morocco on 31 May 2008.

We have also attempted alignments between English and Chinese, but with little success.

We applied distributional factorization to the level-3 features of our adaptive image feature hierarchy without noticing any interesting relationships between the resulting clusterings. Similar results were obtained with low-level VQ codebook features. This was attempted on VOC2007 data and artificial data consisting of oriented rectangles of various sizes. We had hoped to achieve results such as separating scale from shape, but have not yet been successful. However, we found that distributional factorization captures much more information per cluster than ordinary co-clustering, if a Shannon-type (rather than Renyi-type) objective function is used. This shows that the method does find independent degrees of freedom even if their significance is not obvious on inspection. However we also found that this extra information does not significantly improve image classification performance.

3.3 Entity Disambiguation

Improvements were made to our named entity recognition software in order to support an evaluation exercise by SAIC through RDEC experimental center. The main point of the evaluation was to test the software's ability to associate entities across multiple databases with very different schemata. The test problem was to unify entities extracted from FBIS newswire with TIDE-SBU records. The TIDE-SBU data must remain at the UEE lab at Mitre, so a virtual machine was set up there for our use in this project.

Two approaches were taken to handle the disparate schemata. The first, dubbed the "low road" was based on XSLT transforms which had to be written by hand, but made our legacy software automatically applicable to the extent that it handled database fields present in both databases. The other approach, the "high road" required a complete re-write of the system but mostly automated the alignment process. This method assumed only that each database could be dumped to an XML file, and that entity records (which might take the form of sub-trees) were marked by particular tags. Preprocessing was done to flatten the trees and determine and count all the unique data entries in each field. A feature set based on a suite of regular expressions was then derived for each field so that the strings occurring in those fields could be characterized by matches to a modest number of regular expressions. The data structure was then re-expressed in terms of these regular expression features, so that every entity record could be reduced to a feature vector in these terms. These records were matched using a statistical algorithm based on mutual information described in Appendix A.

Satisfactory levels of cross-database entity disambiguation were obtained by both methods.

4 LESSONS LEARNED

4.1 Simple but important

- One must be careful to purge mixed-language documents (English is particularly ubiquitous) from corpora used to train a language ID system.
- Character encoding issues remain a source of difficult bugs in SOA systems due to subtly differing assumptions implicit in different components and programming languages. Other annoying issues plaguing SOA development include JVM version skew, port numbering and firewall problems, and poor debugging diagnostics in service stack programs such as Axis2. Perhaps mixing bleeding-edge software engineering with bleeding-edge AI research is less efficient than a divide-and-conquer strategy.
- Never ask a client or consultant to install software when it is feasible to ship them the hardware with the software installed. This approach was used successfully with the language consultants who labeled data for training NER models, and OSC. The former were sent external hard drives with pre-installed software and data, and the latter were sent a pre-configured server.

4.2 More profound

- One seemingly has to use Renyi mutual information to get distributional factorization to work well, at least for unsupervised language alignment. Then it works well.
- Unsupervised learning of feature hierarchies is feasible for text and imagery.
- Simultaneous adaptation of interest points and feature sets for imagery can be done and produces sensible results.

5 DIRECTIONS FOR FUTURE RESEARCH

5.1 Low priority

5.1.1 Statistical Significance of Mutual Information estimates (long term)

Statistical methods must always confront the issue of statistical significance. The problem may be particularly serious when optimizing over sets of tokenizers that can yield feature sets of very different sizes, and therefore very different statistical estimation error levels for the mutual information quantities in an SDS-based objective function. We applied some effort to the estimation of statistically significant mutual information, following the work of Hutter and Zaffalon in using a Beta distribution to approximate the posterior distribution over mutual information. However we had to drop this line of research in favor of simple expedients based on frequency thresholds in order to concentrate efforts on more pressing concerns.

5.1.2 Hardware acceleration for SDS methods (long term)

We collaborated briefly with Cray on an informal basis to study parallelization of the tokAdapt SDS algorithm developed under NIMD. This is a pure SDS implementation that (slowly) finds semantics-carrying units in text, usually words but often multi-word units or sub-word units. It is slow because the corpus must be re-scanned every time the tokenizer is perturbed. We only got as far as porting the code to a single node of their MTA parallel machine before Cray suspended their effort. There are almost certainly algorithmic optimizations that would help tremendously, and that should probably be pursued first, but SDS inherently involves a large and complex search space, so the issue of parallelization will deserve further attention.

5.2 High priority

5.2.1 Simultaneous unsupervised learning of feature hierarchies

We made enough progress in this project to be confident that Semantically Driven Segmentation (SDS) can be made to work well in the relatively near future (a few years). This is a major enabler of AI, unsupervised learning techniques that convert unstructured information into relational knowledge without relying on detailed prior knowledge of the problem domain or data medium. This is the crux of every information processing problem, and SDS promises a black box that just does it.

5.2.2 Optimization of Distributional Factorization

This project succeeded in producing a highly optimized and effective distributional factorization program and testing it on non-trivial language translation problems, so there is no doubt that the method is feasible and powerful. It was a significant breakthrough to find that a less commonly used definition of information (Renyi information) is most effective, at least for this task. There are several other obvious variants of the objective function that have not been investigated, but should be. Also, further attention to computational efficiency is needed; typical run times are measured in hours.

5.2.3 Application of Distributional Factorizations

This technique ought to be useful for factoring out semantic invariants other than language. In text, it should be investigated not only for more languages, but also for translation by analogy between genres within a language. For imagery, it should be investigated for factoring scale, orientation and lighting from shape. It should be investigated for discovering semantic invariants we never anticipated.

6 TECHNOLOGY TRANSITION STATUS AND ACCOMPLISHMENTS

6.1 Text Analysis Engine

The Text Analysis Engine (TAE) was developed during the NIMD program and further developed under CASE. It is a web service providing access principally to a Named Entity Recognition (NER) capability, but also to entity disambiguation, relationship extraction, document clustering, document cluster naming, and document retrieval functionality. It is part of the **Hydra** system that incorporates services from other NIMD and CASE performers, led by Oculus. Hydra was multiple integration experiments for both the NIMD and CASE programs. The Hydra versions demonstrated how NIMD and CASE technologies could be utilized in an analysts' environment. Figures 6 and 7 illustrate TAE and one of the Hydra integration experiments, respectively.

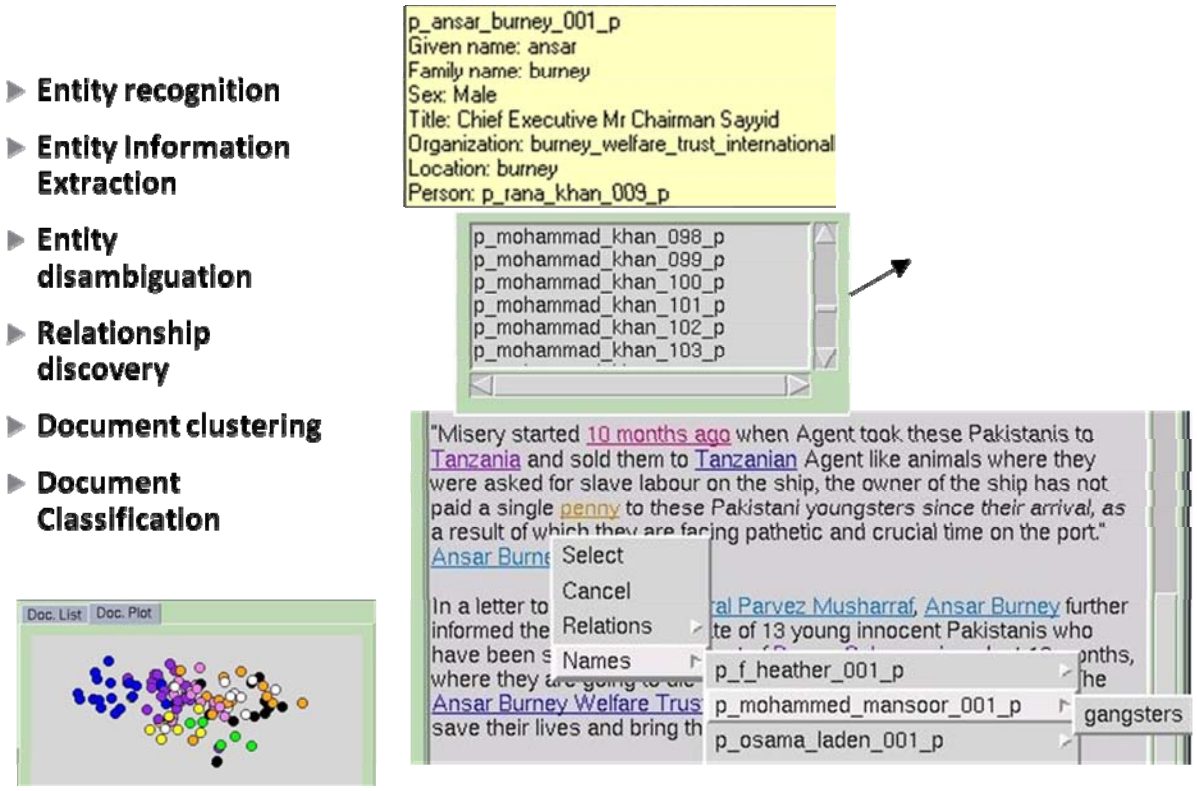


Figure 6: Text Analysis Engine Services (TAE)

We accelerated the document clustering code, in response to complaints from earlier evaluations. The main issue was dealing with long documents, which we simply truncated. Further improvements were made by reducing documents to distributions over term clusters learned from the corpus in an off-line batch run. We made performance and scalability improvements to other important utility programs, including the co-occurrence counter.

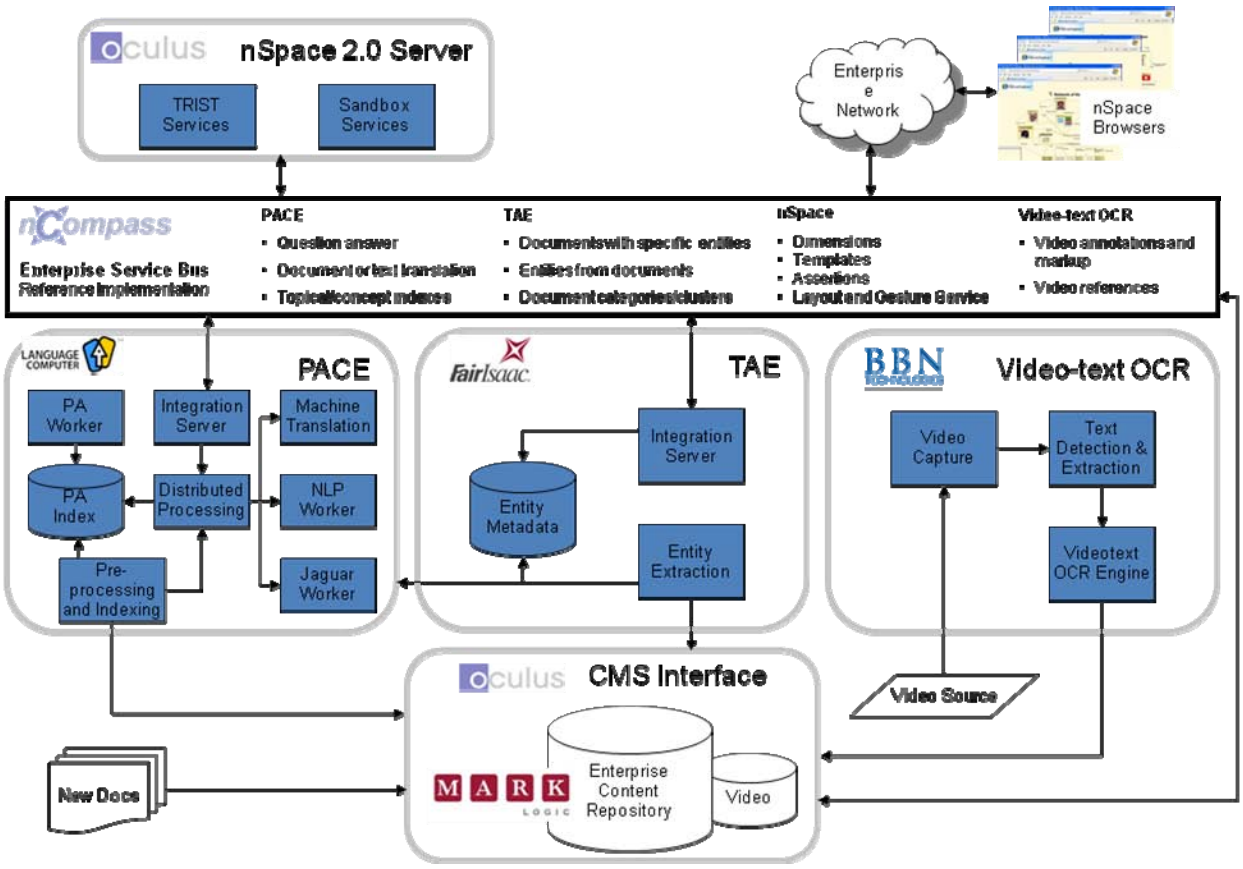


Figure 7: Hydra 3 Architecture

We introduced code for sanitizing wild HTML so the TAE could be directed at data gathered in real time from the Internet.

To support the rapid introduction of new languages (already assisted by unsupervised learning of clusters of semantically related terms), we wrote code for annotating NER data within a browser. Starting from URLs supplied by analysts, we gathered foreign language corpora for training NER models. In Korean, for example, we started from 66 URLs supplied by an OSC analyst, download 1M web pages and culled these to 650K, extracted and tokenized the text, and built clusters for the NER engine. We interacted with our annotators to improve the tokenizers to handle issues such as the writing of prepositions as suffixes in Korean.

We introduced a Bayesian n-gram language ID module to automate switching between NER models. We used foreign language blog data to improve parameter settings, achieving accuracy near 100% on the 8000-document corpus used in the February 2007 exercise at OSC.

The TAE was extended with NER models for **6 languages** (English, Arabic, Korean, Simplified Chinese, Farsi and Russian). With RTTI support added to the CASE project, it was ported from Perl to Java to improve robustness and **scalability**. (Both versions also involve some calls to C routines.) Thread pooling code was added to handle multiple requests. A database abstraction layer was added. The Hydra system was evaluated in February and September 2007. The February exercise used an 8000-document corpus and the July exercise used 20,000 documents. The TAE was also used in the CASE **APEX evaluation** in December 2007. Hydra was also evaluated in August 2008. We used co-clustering on GulfLink Cable data to improve NER for this evaluation. We also used distributional factorization to align Cable data with the web-derived data on which the NER model was previously trained. **RDEC** also did an unanticipated evaluation in August 2007.

Scripts to enable sophisticated users to create their own TAE databases were delivered to Bill Merring of **Open Source Works** (OSW) via Oculus. The TAE can also be populated via web service calls; this is how it is populated from the Oculus CMS and from on-the-fly web search results.

We supplied term clusters to **SET** and **NewVectors** for use in user models in the IE3 demo at the April 2008 PI meeting, and for use in information space models in the **final CASE evaluation** exercise in September 2008.

An API was written for the **distributional factorization** code that should make it easy to incorporate into the TAE. This could be used to provide access to this functionality as a web service in future programs.

6.2 Image retrieval for the Case Integrated Architecture

We worked with Oculus and others in the design and implementation of SOA for handling and processing image data and metadata including attention information. The architecture is based on automated class generation from a WSDL specification layer and XSD configuration files. Our component is an image and/or image-region retrieval service. The SOA was demonstrated at the PI meeting in March 2007 with a mock-up image retrieval service and a Helipad recognition problem. It is currently undergoing modification and incorporation of a functional image retrieval service (albeit in its current, truncated state of development) for the final **HILATR evaluation** exercise to measure the improvement of SAM site recognition with the introduction of human interest data based on eye tracking and EEG measurements.

7 ACRONYMS AND ABBREVIATIONS

- **AFRL** Air Force Research Labs.
- **AGS** Association-Grounded Semantics.
- **AI** Artificial Intelligence.
- **CASE** Computer and Analyst/System Effectiveness.
- **DA** Distributional Alignment.
- **DTO** Disruptive Technology Office.
- **EEG** Electro-Encephalogram.
- **FBIS** Foreign Broadcast Information Service.
- **HILATR** Human-In-the-Loop Automatic Target Recognition.
- **HNC** Hecht-Nielsen Neurocomputing Corporation
- **HTML** Hypertext Markup Language.
- **IARPA** Intelligence Advanced Research Projects Activity.
- **JVM** Java Virtual Machine.
- **NER** Named Entity Recognition.
- **NIMD** Novel Intelligence from Massive Data.
- **OSC** Open Source Center.
- **OSW** Open Source Works.
- **RESOUND** Reasoning Efficiently from Self-Organization of UNstructured Data
- **RDEC** Research and Development Experimental Collaboration.
- **RTTI** Rapid Technology Transfer Initiative.
- **SAIC** Science Applications International Corporation.
- **SAM** Surface-to-Air Missile.
- **SDS** Semantically-Driven Segmentation.
- **SOA** Services-Oriented Architecture.
- **SRI** Stanford Research Institute
- **SVM** Support Vector Machine.
- **TAE** Text Analysis Engine.
- **TIDE-SBU** Terrorist Identities Datamart Environment – Sensitive But Unclassified
- **URL** Universal Resource Locator.
- **VQ** Vector Quantization.
- **WSDL** Web Service Definition Language.
- **XML** Extensible Markup Language.
- **XSD.** XML Schema Definition.
- **XSLT** Extensible Stylesheet Transformations.

APPENDIX A – STATISTICAL ALGORITHM

This appendix describes the statistical algorithm used to match feature vectors for entity disambiguation.

The distributional clustering framework

Entity disambiguation can be regarded as a clustering problem in which all *instances* or *mentions* of an entity are grouped according to which instances refer to the same entity.

Each instance of an entity can be converted into a set of features values. The features can include the textual strings representing the entity names, features such as regular expressions that match those strings, and features derived from the context of the instance. In a database record, that context is the content of the other fields in the record besides the entity. In free text, the context is the text surrounding the string that directly represents the entity.

Although the string representing the entity name is intended to be a direct indicator of the entity, and the context is generally regarded as ancillary evidence for or against particular entities, there is no essential mathematical distinction between the name and its context. The string that names the entity is merely a particularly strong form of evidence (one hopes) for a particular entity, and the context is generally weaker yet helpful evidence. Because there is no fundamental distinction between these forms of evidence, our formalism does not prejudice the analysis by introducing one. Features derived from (purported) names have the same status as features derived from the context. Aside from conferring a theoretical advantage, this helps to circumvent difficulties with introducing arbitrary rules about distinguishing a name from its context, such as rules as to whether titles and honorifics are to be regarded as part of a name.

For us, then, an instance of a named entity is a set of feature values. To simplify the formalism, we take these features to have binary values, so we can say that a feature is ``present" or ``absent" from a mention. This involves no loss of generality because features with more than two possible values, such as most string-valued features, can be reduced to a set of binary features by declaring every possible string to be a feature, the value of which, in any particular mention, is a Boolean indicator value. In our system we use a more carefully selected set of string features, each of which is defined by a particular regular expression. We use the notation $\phi_i^{(t)}$ to designate the Boolean value of feature *i* in the $t^{\underline{\text{th}}}$ instance an entity in some standard numbering of all the entity instances in the data set concerned. The set of features comprising mention *t* is denoted $\phi^{(t)} = \{\phi_1^{(t)}, ..., \phi_n^{(t)}\}$, where *n* is the total number of features.

Maximization of mutual information is a theoretically sound and experimentally well-tested clustering methodology. In the present case, we introduce the joint random variable (Φ, H), any instance (ϕ, h) of which is a set of *n* Boolean feature values $\phi = \{\phi_1, ..., \phi_n\}$ together with an entity *h*. Any particular instance occurs with some probability $P(\phi, h)$, and from this density we can obtain marginal and conditional probabilities in the usual way. The mutual information between the features and the entities is

$$I_{\Phi H} = \sum_h P(h) \sum_\phi P(\phi|h) \ln \frac{P(\phi|h)}{P(\phi)} \qquad (1)$$

Recall that each entity *h* is a set of instances, and the overall problem is to determine these sets. We will proceed by starting with singleton sets and progressively merging them, determining whether a merger is warranted according to the impact it would have on $I_{\Phi H}$. This impact will be low if the feature sets of the merge candidates are very similar. The bulk of the problem is to devise a practical yet plausible approximation to the change in $I_{\Phi H}$ when hypothesized sets of instances are merged.

Approximating the change in mutual information

The mutual information is a property of the joint distribution over the possible values of all the features and the entity. Given a reasonably large number of features, most of the possible combinations of features will never occur in a data set of reasonable size, so it is not possible to estimate this joint distribution directly. Therefore we will employ an estimate based on combinations of marginal and conditional distributions over small enough sets of features to give significant results, choosing this combination with a view not only to practical considerations, but also to how the approximations impact the estimate of mutual information.

The joint probability $P(\phi, h)$ can always be factored into the form

$$P(\phi, h) = P(h) \prod_{i=1}^{n} P(\phi_{x_i} | \phi_{x_{i-1}}, ..., \phi_{x_1}, h). \qquad (2)$$

Using a factorization of this type in the numerator and denominator of the logarithm in (1) gives

$$\begin{aligned} I_{\Phi H} &= \sum_h P(h) \sum_{\phi_{x_n} ... \phi_{x_1}} P(\phi_{x_n}, ..., \phi_{x_1}|h) \sum_i \ln \frac{P(\phi_{x_i}|\phi_{x_{i-1}}, ..., \phi_{x_1}, h)}{P(\phi_{x_i}|\phi_{x_{i-1}}, ..., \phi_{x_1})} \\ &= \sum_h P(h) \sum_i \sum_{\phi_{x_i} ... \phi_{x_1}} P(\phi_{x_i}, ..., \phi_{x_1}|h) \ln \frac{P(\phi_{x_i}|\phi_{x_{i-1}}, ..., \phi_{x_1}, h)}{P(\phi_{x_i}|\phi_{x_{i-1}}, ..., \phi_{x_1})} \qquad (3) \end{aligned}$$

where we have used normalization to sum out features not present within the logarithm.

In order to collapse the nested sum over features, we replace $P(\phi_{x_i},...,\phi_{x_1}|h)$ by its maximum likelihood approximation in the term outside the logarithm in (3). The maximum likelihood estimate of $P(\phi_{x_i},\phi_{x_{i-1}},...,\phi_{x_1}|h)$ formed from N_h mentions

$$(\phi_{x_1}^{(1)},...,\phi_{x_n}^{(1)}),...,(\phi_{x_1}^{(N_h)},...,\phi_{x_n}^{(N_h)}) \tag{4}$$

is

$$P(\phi_{x_i},\phi_{x_{i-1}},...,\phi_{x_1}|h) \approx \frac{1}{N_h}\sum_{t=1}^{N_h}\prod_{j=1}^{i}\delta_{\phi_{x_j}\phi_{x_j}^{(t)}} \tag{5}$$

The maximum likelihood estimate is a poor one, but more sophistication seems unlikely to yield much improvement when N_h is very small. Making this substitution gives

$$I_{\Phi H} \approx \frac{1}{N_h}\sum_h P(h)\sum_{t=1}^{N_h}\sum_i \ln\frac{P(\phi_{x_i}^{(t)}|\phi_{x_{i-1}}^{(t)},...,\phi_{x_1}^{(t)},h)}{P(\phi_{x_i}^{(t)}|\phi_{x_{i-1}}^{(t)},...,\phi_{x_1}^{(t)})} \tag{6}$$

The same substitution could have also been made in the numerator of (3), but we choose not to do this at this stage. We note, however, that the sums are collapsed onto precisely those partial feature vectors $(\phi_{x_i}^{(t)},,...,\phi_{x_1}^{(t)})$ for which non-zero number counts are available to estimate $P(\phi_{x_i}|\phi_{x_{i-1}},...,\phi_{x_1}|h)$, so the maximum likelihood estimate of $P(\phi_{x_i}^{(t)}|\phi_{x_{i-1}}^{(t)},...,\phi_{x_1}^{(t)},h)$ in this numerator is never 0.

The effect on $I_{\Phi H}$ of creating an entity $h+h'$ by merging *h* with *h'* is

$$N\Delta I_{\Phi H}(h,h') \approx \sum_i\left[\sum_{t=1}^{N_h+N_{h'}}\ln\frac{P(\phi_{x_i}^{(t)}|\phi_{x_{i-1}}^{(t)},...,\phi_{x_1}^{(t)},h+h')}{P(\phi_{x_i}^{(t)}|\phi_{x_{i-1}}^{(t)},...,\phi_{x_1}^{(t)})}\right.$$
$$\left.-\sum_{t=1}^{N_h}\ln\frac{P(\phi_{x_i}^{(t)}|\phi_{x_{i-1}}^{(t)},...,\phi_{x_1}^{(t)},h)}{P(\phi_{x_i}^{(t)}|\phi_{x_{i-1}}^{(t)},...,\phi_{x_1}^{(t)})}-\sum_{t=1}^{N_{h'}}\ln\frac{P(\phi_{x_i}^{(t)}|\phi_{x_{i-1}}^{(t)},...,\phi_{x_1}^{(t)},h')}{P(\phi_{x_i}^{(t)}|\phi_{x_{i-1}}^{(t)},...,\phi_{x_1}^{(t)})}\right] \tag{7}$$

Here we have used the maximum likelihood estimate $P(h)=\frac{N_h}{N}$ to replace $\frac{P(h)}{N_h}$ with $\frac{1}{N}$.

By defining the ``score"

$$s_h = \sum_{t=1}^{N_h}\sum_i \ln\frac{P(\phi_{x_i}^{(t)}|\phi_{x_{i-1}}^{(t)},...,\phi_{x_1}^{(t)},h)}{P(\phi_{x_i}^{(t)}|\phi_{x_{i-1}}^{(t)},...,\phi_{x_1}^{(t)})} \tag{8}$$

(7) can be read as the score of the merged entity less the scores of the entities that were merged. We can therefore focus on the calculation of the score (8). It is easily shown that mutual information can only be decreased by merging categories, and that (8) is non-negative, so the score of the merged entity will never exceed the sum of the scores of the entities that were merged. The question arises as to how to decide a threshold for $|\Delta I_{\Phi H}(h,h')|$, above which merges should cease. This can be decide by collecting the values of $|\Delta I_{\Phi H}(h,h')|$ as the merging proceeds, sorting them, and looking for derivative discontinuities in the resulting curve.

There is an asymmetry in the uncertainty we should attach to probability estimates based on zero counts, particularly in the extreme case $N_h=1$. If $N_{1|xh}=1$, this can only be due to $\phi_x^{(t)}=1$ in the data for *h*. While we could reasonably presume the maximum likelihood estimate $P(\phi_x|h)=1$ to be too high, we could be fairly sure that $P(\phi_x|h)>>0$, because we expect most features to be absent from most mentions, so the presence of a feature is more likely due to a probability well removed from 0 than a statistical accident with a

nearly vanishing probability. The same cannot be said for the case $\phi_x^{(t)} = 0$. Because typical mentions have few features present, and it is common for different features to be present in different mentions, it is entirely plausible in this case that $P(\phi_x|h) >> 0$ but feature *x* just happens to be absent in the single mention on hand. Therefore we attach higher uncertainty to small probability estimates than to large ones.

For this reason, the likelihood ratio in (8) is highly uncertain in the case that $\phi_x^{(t)} = 0$; the numerator in particular is uncertain, and a reasonable guess for its value is the background probability in the denominator. This provides some justification for discarding these terms in the sum on *i*. We can go further. We have already argued that the numerator will be reasonably large because the probability is evaluated on a data point used to estimate it. This applies regardless of the value of $\phi_x^{(t)}$. But we also expect the denominator to be fairly large just because it is evaluated on $\phi_x^{(t)} = 0$, and we expect the background probability for absence to be high for most if not all features. Therefore we will neglect these terms in (8) which becomes

$$s_h \approx \sum_{t=1}^{N_h} \sum_i \delta_{\phi_{x_i}^{(t)},1} \ln \frac{P(\phi_{x_i}^{(t)}|\phi_{x_{i-1}}^{(t)}, ..., \phi_{x_1}^{(t)}, h)}{P(\phi_{x_i}^{(t)}|\phi_{x_{i-1}}^{(t)}, ..., \phi_{x_1}^{(t)})}. \qquad (9)$$

In this way the score for each entity hypothesis becomes a sum over only those features which occur in its instances.

As a further approximation, we merge all the data for an entity into a single ``virtual" mention, so the sum over the N_h mentions is replaced by a factor of N_h. This makes the numerators better estimated, statistically, at the expense of losing some within-entity mention-level correlation structure. The score then becomes

$$s_h \approx N_h \sum_i \delta_{\phi_{x_i},1} \ln \frac{P(\phi_{x_i}|\phi_{x_{i-1}}, ..., \phi_{x_1}, h)}{P(\phi_{x_i}|\phi_{x_{i-1}}, ..., \phi_{x_1})} \qquad (10)$$

in which we no longer need to distinguish mentions *t*.

Next we turn our attention to choosing an ordering for the features in (10), and workable approximations for the likelihood ratios.

Aside from the factor N_h, the first term of sum (10) is $\ln \frac{P(\Phi_{x_1}=1|h)}{P(\Phi_{x_1}=1)}$ for whichever feature we choose for x_1 amongst those that are present in *h*. With maximum likelihood estimation, the numerator will be 1.0 if $N_h = 1$, and at least $\frac{1}{N_h} >> 0.0$ for larger N_h, so this term will contribute roughly the logarithm of the background probability of x_1.

For the reasons discussed earlier, we can consider all the features for which $\phi_x = 1$ to be likely for entity *h*. Conditioning on the presence of more features that are characteristic of *h* should not change that situation much. Therefore we can assert $P(\phi_{x_i}|\phi_{x_{i-1}}, ..., \phi_{x_1}, h) \approx P(\phi_{x_i}|h)$. For the background, we expect features to come in overlapping cliques that tend to be mutually predictive, due to a tendency to appear together in several different entities. We want a procedure that approximates counting each clique only once, because otherwise it would be possible to scale up a term of (10) by adding redundant features.

As one proceeds from term *i* to term *i+1* in the sum (10), feature ϕ_{x_i} is moved from the left of the conditioning bar to the right, and a new feature $\phi_{x_{i+1}}$, selected from those remaining, is placed on the left. We consider the rarest (most predictive) feature first, thinking that these are likely to be the most predictive of the most other features. We want to condition on the highly predictive features as early as possible because these ``explain" the remaining features in their ``clique", in the sense of giving a high value of $P(\phi_{x_i}|\phi_{x_{i-1}},...,\phi_{x_1})$ and therefore a small contribution to the score. In this way, the first feature of a clique that moves across the bar makes a large contribution, but the remaining features of the clique do not. We also approximate

$$P(\phi_{x_i}|\phi_{x_{i-1}},...,\phi_{x_1}) \approx \max_{j<i} P(\phi_{x_i}|\phi_{x_j}) \tag{11}$$

conditioning only on the feature x_j that makes x_i most probable. A more complex variant

$$\max_{j<i}\left(p_{x_i|x_j} + \max_{k<j} p_{x_i|x_k}(1-p_{x_k|x_j})\right) \tag{12}$$

attempts to account for conditioning features in addition to x_j by adding the probability conditioned on the feature least well explained by x_j.