Analogy in the Large

Kenneth B Haase
MIT Media Laboratory
20 Ames Street, E15@85
Cambridge, Massachusetts 02142
USA
haase @ media mit edu

Abstract

This article discusses the use of analogy to index and organize large databases of information We describe the design and implementation of an analogical database supporting tens to hundreds of thousands of cases The contents of the database are parsed news articles represented as networks of grammatical relations with references into WordNet for word meaning information The virtue of this approach is its domain independent handling of content analysis Efficient algorithms for indexing and matching in this database are described and bneflv discussed and examples of their performance are discussed

1 Introduction

This article discusses the design of databases which support dynamic analogies among thousands of descriptions We have built an analogical database consisting of (currently) over a million words of natural language text parsed into networks describing surface syntactic structure and associated with a global ontology derived from WordNet [Miller 1990] Queries to the database retrieve networks with similar grammatical structure and match elements in order to identify thematic roles We call the database analogical because the mappings between elements are not determined by *a prion* canonical structures like case frames scripts or memory packets but by drawing analogies between them on the fly The technical contributions of this work include algorithms for efficiently determining analogies an analysis of the problems of indexing for analogy among thousands of descriptions and an implemented indexing system based on the analysis In addition we discuss the use of Wordnet as a semantic background for analogizing and indexing

2 The Database

We are currently constructing a database of over 10 000 000 words of parsed text for use in experiments on domain-independent text analysis The text corpus was provided by Gannett Corporation and consists of a large number (> 50 000) of relatively short (usually one or two paragraph) news summaries drawn from the periodical *USA Today,* and ranging from 1987 to the present

The database itself (of which roughly 10% has been processed as of April 1995) is generated by parsing the text into networks of frames representing individual phrases and interconnected by links reflecting possible grammatical relations among them Individual frames are connected with an ontology of possible meanings derived from WordNet

The chief virtue of the database (and our approach) is that it allows processing based on semantic content without the need lor either the design of canonical meaning representations or the implementation of processes for producing them

3 Representational Structure

Descriptions in our database consist of sets of individuals connected by two kinds of relations

- micro-relations connect individuals *to* either other individuals in the same description or literal values
- associations connect individuals to either individuals in other descriptions or to reference points in the global namespace

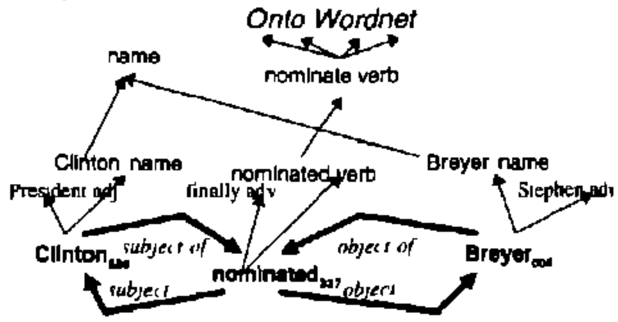
Micro-relations provide *a pre semantic structural representation* which is neither canonical (there is no promise that semantically equivalent descriptions have the same micro-relational structure) nor entirely correct (some micro-relations may be accidental or artifactual) Micro-relations constrain but do not constitute interpretations

Associations provide an *ambiguous meaning representation* such that if two individuals have associations in common, they are taken as having some semantic similarity Like micro-relations associations may be ambiguous and partially incorrect The association relation is transitive (if x is associated to v and v to z then x is associated to z) but usually not symmetric (il x is associated to y y is not neccessanly associated lo x) The immediate associations of an individual constitute a set out of which all of its associations can be generated

Micro-relations and associations are central to the matching and indexing processes the matcher uses associations! structure to provide base-level matching and micro-relational structure to determine higher level matches The indexer combines associational and micro-relational structure to construct signatures for descriptions such that overlapping signatures are indicative of systematic and sensible analogies

3 1 Representing Text

For example, in our text database a sentence like President Clinton finally nominated Stephen Breyer is translated into a description



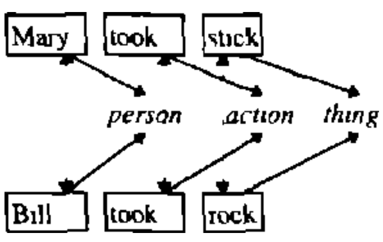
consisting of three individuals (in bold face) four micro relations (heavy lines) and a number assoications with 'global reference points including entries in WordNet

4 Base Level Matching

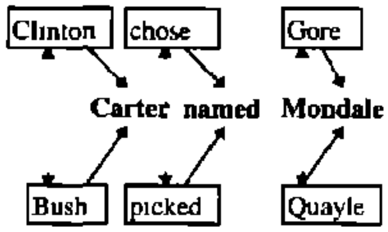
The chief innovation of our matcher is its use of a network of associations to determine base-level similarity Analogical matchers like SME [Falkenhamer *et al* 19891 have a basic and fixed level of symbolic description at which the matching process grounds out In our matcher the base level consists of a network of associations which grows as new descriptions are indexed and matched Two important properties of this metric are that the criteria of similarity are (a) contextually sensitive and (b> can be changed without changing the implementation or the matcher A third feature is that the introduction of new elements to the base level occurs naturally with the accumulation ot new cases and their association with existing cases This second feature distinguishes it from the ACME matcher [Holyoak and Thagard 1989] which uses an associational base level but implements it with a fixed network

The base level similarity metric for our matching mechanism applies to the individuals which constitute descriptions and is based on the identification of unique common associations between them Two individuals are cognates (i e similar with respect to their contexts) if they share some unique common association *z* which is shared by no other pairs of individuals from the two descriptions The cognate relation has two interesting special cases type unique cognates and triangle cognates

Type-unique cognates occur when two heterogenous descriptions are compared and the associations which make the elements distinct become the foundation for cognate relations between the description For example in the following network the unique associations for people actions and things sort out the cognates between two situations



Triangle cognates occur when common associations are individuals in a third description If two descriptions already have a set of associations in a third description these can provide the basis tor cognate relations between them For example



Triangle cognates obviate abstract schemata for descriptions by allowing one concrete example to provide structure to others and by permitting past associations and analogizing to support current and future matching

Of course when presented with a pair of descriptions whose common associations are very general cognate matching will produce fanciful results However the goal of this base level of matching is not to generate only *sensible mappings but to generate the most sensible mapping given the possibilities*

4 1 Computing Cognates

Given two descriptions C and D the cognate relations between them can be determined in *0(mn)* time where n is the size (number ot elements) in C and D and m is the depth of the association tree tor the elements

In our text database with the association network derived from Wordnet *m* ranges trom 2 to SO with an average value of about 10 This means that — given the appropriate associauons analogical retrieval of description components can be done in *0(m)* lime for each component, performance which allows us to use analogy as the basis for providing representational structure

The cognate determination algorithm is a two-phase competitive algorithm where members of one context first compete to uniquely mark their associations and then members of the other context compete to claim the associations which have been uniquely marked Whenever a conflict occurs in either phase the common association drops out of the running

By using bits and a tag field on each description each phase takes *0(mn)* time and a final cleanup phase takes *Oin)* time, giving *Ctftnn)* time overall

Cognate matching provides a base matching level for analogy which is botfi flexible and efficient It also in the case of triangle cognates allows the reuse of associations determined in the past to generate new mappings essentially memoizing past analogical work Cognate matching will fail to match two elements if there are either

1 *no* common associations between them
2 no *unique* common associauons between them

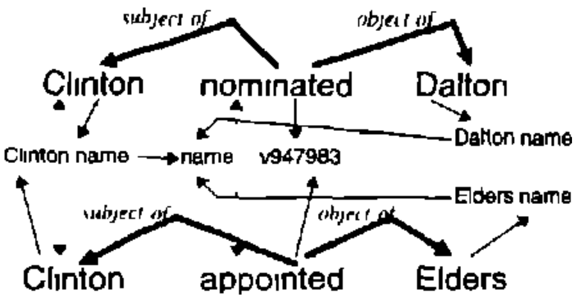
Case (1) is relatively uncommon when the database has a rich associational structure (such ^*s* Wordnet) Case (2) is more common and we resolve it though the use of structure matching to combine associations and micro-relations to resolve ambiguities

5 Structural Matching

Structure matching in our database starts with a set of *initial mappings* denved by cognate matching and extends this mapping based on the micro-relational structure of the description Unlike SME, our algorithm does not explicitly represent spaces of possible mappings instead *it* generates a single map which will be ambiguous if there is no one systematic mapping

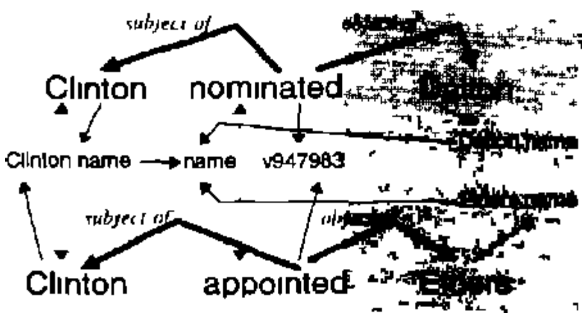
The algorithm starts from an initial set of pairings determined on the basis oi cognate relations, it then considers each pair and uses the micro relations of the paired individuals to specify smaller sets out of which it attempts to resolve cognates

For instance in matching the following networks we can determine two cognate relations



the relation between named and appointed is found through common associations in Wordnet (v947981) the relation between the two Clintons is based on their common root (namely Clinton name') However no mapping between Dal ton and Elders can be found because its common association (name) is also common with Clinton

Structure matching however combines the mapping between nominated and appointed' and the micro-relations object of to select a subset to compare for cognates



Restricted to this subset the conflict with Clinton does not exist and the mapping between 'Dalton ' and "Elders is easily generated

An $G(tnn^2)$ algorithm for structure matching starts by generating the cognate mappings and then expanding each pairing Expanding a pairing involves iterating over the common micro-relations for each and computing the cognates between the sets ot frames to which they are (respectively) micm-reJated

There will be $0(n)$ matches to expand at most each expansion will take $0(kmn_k)$ where k is the number of micro-relations and n_k the number *of* elements to which each element is connected by each micro-relation In practice k is usually small (< 10) and n_k is bounded by n This gives bounds of $0\{nuC)$ Lhough in many cases (including the representation of grammatical/semantic structure) n_k tends to be small (< S), giving bounds more like the same $0(mn)$ required for cognate matching

Because this algorithm relies on cognate relations as the base level for matching it is prone to the same sort of fanciful results as cognate matching For instance a description of two people kissing one another and two people hitting one another would be matched based oti the fact that both actions have a common synsel (roughly concept) in Wordnet (e g contact or touch) But as with cognate matching the goal of matching is not to produce only sensible matches to produce the most sensible match given the possibilities The task ot maintaining sensibility falls on the coverage of the database and the indexing mechanisms which retrieve descriptions for possible matching Tt is 10 these components ot the database that we now turn

6 What Makes a Good Match[9]

How do we decide what makes a good match for a description[7] How do we find such matches without trying to match with every description in the database" In this section we describe how our databases arc indexed starting with a characterization of what makes a good match and then describing the indexing scheme for identifying such matches wilhout examining each description individually

We define a good match to have two interrelated features systematicuy and solidity The concern lor systematicity is common to nearly all work on analogy [Falkenhainer *et al* 19891 [Holyoak and Thagard 19891 and [Mitchell 1991] all seek systematicily albeit in different ways

But as we saw in the kiss/hit case our structural matcher is perfectly happy to generate fanciful and unreasonable systematic matches To address this problem we add the criteria of match solidity to match systematicity in our evaluation of matches The solidity ot a particular mapping is a function ol the number of unique common associations which support it in the case ol the fanciful match between Bill hit the table and Bill painted the flower the match is supported by a slender thread ot associabon through the Wordnet svnsets tor actions and things In contrast a match between Bill hit table and Bill kicked the ball would be supported bv many more unique common associations

Solidity is related to the stability ol mappings when extraneous elements are added in descriptions *If* we extended Bill hit the table to be Bill grabbed the table and hit it the match to Bill painted the flower would be lost because the common association between hit md painted is no longer unique with another action to compete with it On the other hand the match hetween Bill grabbed the table and nil *it* and Bill kicked the bill would survive the addition ot the distracting elements

7 Indexing for Good Matches

Given this informal account of systematicity and solidity we move on to the question of how we use the criteria of systematicity and solidity in constructing an index for analogical matching

We start with the fact that a match is *systematic* if corresponding elements are related to one another by the same micro-relations and that match is *solid* if the corresponding elements also have a lot of unique common associations Uniqueness however is a contextual property which requires looking at the descriptions being matched Because of this our signature must be based only on common associations and not on unique common associations

To capture these dual constraints we define a description s signature as a set of distinct keys such that overlap of signature indicates the potential for solid and systematic matches The general form of a key is a triple of an individual s association a micro relation and an association of the related individual For instance some of the keys for the Clinton said' example above might be

<Clinton,subject-of,said>
<person,subject-of,said>
<Clinton,subject-of,communicated>
<person,subject-of,communicated>

The full signature of a description just includes all the associations and micro-relations of a description e g the set of 3-tuples

$$\{\forall x \in C \ \ r \in R_x \ \ A(x) \times r \times A(r(x)) \}$$

where C is a description R_T its micro relations *A(x)* the associations of *x* and *r(x)* the elements of *C* to which *x* is micro-related by *r*

The connection between the overlap of signatures and svstematicity and solidity goes as follows For every systematic relation carried over in a mapping between descriptions C and D there will be at least one common kev in K If the systematic relation is supported by more common associations there will be more keys in common

Sharing a key constitutes a neccessary but not sufficient condition for a systematic mapping The sufficient condition is a contextual one whether or not the common associations described by the keys are in fact unique given the two descriptions being matched

The chief problem with this approach is that the size of the signature lor a description is a function of the square of the depth of the association network times the number of elements and micro-relations in the description This amounts to several thousand keys for even small descriptions For practical purposes a is neccessary to index with less than a full signature and the identification of this reduced signature is an open research problem We are currently indexing words based soley on stemming (e g said goes to say) and then expanding queries at search time to include direct synonyms

Our use of flat indexing is similar to that of ARCS [Thagaid 1990] and MAC/FAC [Gentner *et al* 1991] but differs in using a key space which reflects relational as well as associational structure The current approach is also similar to *keyword expansion* whose precision problems are described in [Voorhees, 1994] however the addition of structural information allows us to handle some of this loss of precision by rejecting matches based on structural context and syslematicity

The problem of indexing these superfical descriptions is still an open one Other possibilities we a*re* currently examining include

1 Statistical analysis to determine independence and significance of keys
2 Selection of basic types in the Wordnet association network to use as keys
3 Disambiguation of word sense to reduce the overall signature

8 Generating the database

This section describes the generation of the database from input text The text database is generated from input text in a four phase process

1 Tagging breaks the document into words and determines parts of speech using a hand-coded probabalistic grammar which demonstrates 96% accuracy when run (without specialized training) on the Brown corpus
2 Phrasing identifies atomic phrases in the text and their heads the tag set is subcategonzed to support effective phrasing
1 Grounding creates new frames tor head nouns and verbs in the document and associates these frames to frames in a global database and through there to a transcription of Wordnet into the frame database
4 Linking hooks up the local frames for a document based on possible grammatical relations between the phrases they describe Linking is done by a suite of specialized procedures (rather than a general grammar as in [Sleator and Temperly 1989])

The parser operates at roughly 2 000 words/min when running on a single machine the tagging and phrasing is done by a Lisp program while the grounding and linking takes place in Scheme The modules communicate via a LISP-based remote procedure call protocol

Interested readers can experiment with the parser and text matcher al the World Wide Web site

`http //parser media mit edu/demos/`

Note that the goal of the parser is not to produce any one interpretation of the text but to generate a set of structures which will constrain and guide indexing and matching The process is intentionally over-generative in producing multiple attachments for prepositions and ambiguous sense references It is the task of the matcher and database to sort out these ambiguities

9 Performance

When phrase structure and word choice is very stylized and similar between texts cognate and structural matching usually has an easy time determining correspondences between texts For instance in the daily market reports included in the database, the following texts (typography indicates different match derivations) were easily matched

J The Dow Jones average of 30 industrials opens at a record *3734 53* Thursday after closing up 15 6S Wednesday The N4SD4Q OTC composite opens at 767 *89,* down *1 46*

2 The Dow Jones average of 30 Industrials open?, at 3685 4 Friday after closing down 19 0] Thursday The NASDAQ OTC composite opens at *754 14* down 8*02*

Bold words were immediately identified as cognates underlined words *are* matched based on micro-relations between cognates *and italuized words* (the numeric and temporal particulars) are matched based on micro-relations between the bold and underlined words The process easily aligns the corresponding numbers reported in the two texts

In this particular case the chief virtue of our matcher is its ability to operate on the text without priming While it would be straightforward to construct a regular expression to extract that particular numeric value from thai particular class of d nly report the structures and algorithms we have described do so directly without any external intervention

9 1 Harder Matches

Of course the chief reason for the easy success in the above ease was that its phrasing and wording were highly stylized Among the goals tor our past year of work were the expansion ot automatic matching to less stylized cases through a combination of some phrasal canonicalization (retaining ambiguity) and the use of WordNet to represent knowledge about word meanings In this example we collected various reports on administration appointments and produced representations lor them which were then matched The results were satisfying The following four sentences despite differences in wording and spelling can have many of their thematic elements mapped to one another automatically without any introduction of special representations or encodings

1 President Clinton nominated, outspoken Jocelyn Elders-) to be his surgeon general on Thursday

2 Clinton named William Perry deputy secretary under Aspin to the post

3 President Clintoni fired embattled FBI $chief_{l0}$ William Sessions Monday and is ready to nominate, louis Freeh a federal judge in New York City and former FBI agent

4 ' San Anlomo banker John Daltong a former Navy submarine officer^ and Democratic fund- raiser-; was $chosen_8$ Wednesday by President Clinton to be Navy secretary

In sentences 1+2, the phrase structure is exactly the same, and the connection through Wordnet handles the variation in word choice In sentence 3 a policy of projecting subjects forward to capture embedded clauses (when there is not a conflicting subject) connects the President Clintoni firing Sessions to the expected $nomination_4$ ot Freeh In sentence 4 the rules for transforming the passive sorts out subject and object, allowing $Dalton_5$ to match the corresponding elements of the other sentences

Sentence 4 also demonstrates the advantage of representing ambiguity explicitly The sentence s representation is explicitly ambiguous about which phrase $Dallon_5$, submarine $officer_6$ or $fund\text{-}raiser_7$ is the actual subject of $chosen_8$ However when asked to determine a match with a particular second text the unique common prototype relation pulls up the person s name in one relational context as analogous lo the person s name in the other But the representation of matching allows us to both simplify the parsing process *(by* postponing resolution) and simplify the matching process (by not having to consider parser errors)

While the system can do a pretty good job ot figuring out who was nominated only in sentences 1 and 4 is it able to figure out what position thev were being nominated to Fill It misses out on 2 and 3 tor two different reasons

- For sentence 2 it does not resolve the anaphoric referent of the post Solving this requires some mechanism lor intersentential anaphora we do not currently have one but expect thatwe will be able to take advantage of the same representation of ambiguity used tor sense and attachment lo represent a space of possible referents
- For sentence 3 the reason for the system s ignorance is the common sense or conventional inference an astute reader makes that if the sentence mentions someone being fired from a position and then describes a planned nomination that it s likely to be a nomination to that position

There can ot course be no general solution to the problem of conventional knowledge required for sentence 3 since it is contingent and cultural by its very nature *One* possible way to allow the system lo accquire this kind of knowledge would be a framework where new sentences were automatically associated -- upon arrival — with existing sentences having similar structure Different structures with similar meaning could then be associated with one another and through these common associations the new sentences would likewise be associated with each other despite the differences in phrasal conventions

For instance if sentences 1 and 3 were aligned to demonstrate meaning equivalence (e g surgeon general associated with FBI $chiet_{10}$) and a subsequent sentence arrived

5 President Clinton fired embattled surgeon $general_{10}$ Jocelyn Elders and is ready to nominate

Its description would be aligned with Sentence 3 above and by common association generaln would be cognates with chiefs as well as the corresponding elements of other texts associated with Sentence 3

Of course it remains to be seen whether this approach to accquinng this sort of knowledge is effective it might he that either variations are too large or their natural structure too contusing to allow this approach to succeed One of our

Subscripts on the words are used to distinguish particular word occurrences for discussion in the text hopes is that the historical breadth of the database (7 or 8 years of news) will allow us to provide examples of phrasal variation over one period of time and then examine how well those examples cover other periods

92 Indexing Performance

Results on indexing performance on our current database are still preliminary but some interesting problems have emerged When indexing on literal word roots (e g 'met' matches only met) retrieval usually manages to identify texts with similar meaning e g for Clinton met Mitterand" the system found sentences like

> Chinese President Jiang Zemin will meet Russian President Boris Yeltsin Sept _2 in Moscow

but also made the understandable confusion

> The USS Brewton met the Hokulea about 550 miles southeast of Hawaii and picked up

and in both cases identified the active subjects and objects by simple structure matching

For some cases the synset-based expansion is quite successful lor matches to 'Police arrested Simpson texts such as

> CAUGHT Fred Hamilton 34 was captured in Hinion Okla a week after he and two other murderers

are readily retrieved Analogical matching here succeeds in extracting 'Hamilton as analogous to Simpson However overall matching only got 54% of the arrested individuals But a closer examination revealed that most of the misses were due to confusions that categonzcd places or days of the week as individuals a deficit currently being corrected

In addition when WordNet is used to expand the query problems sometimes emerge because individual words are not disambiguated and different senses collide Thus a search for Clinton met Mitterand misidentified

> SPACE STATION NASA said it cant meet President Clintons goal of building a space station for $9_billion

because it confuses meeting a goal with meeting a person However this occurs with a lower score since the object relation of 'meet to a person does not exist Unfortunately there is no such lower score exists with the retrieval of

> Among Tour parades Sunday King of Bacchus the Greek wine god this year played by martial arts film star Jean Claude Van Damme

based on the synset (tor athletic competition) containing both 'played and met neither of which is apropos

These are similar to the problems described in [Vorhees 19941 with keyword expansion we are currently considering planning to use sense ordering information available in the latest version (15) of WordNet to ameliorate some of the problems Another possible solution is to try some automatic clustering on the document level hoping that additional context in actual articles may resolve some of the ambiguity which is causing our problems Another more labor intensive approach is to use a corpus of text which has already been disambiguated as a corpus against which new texts are indexed

10 Ongoing and Future Work

We currently have only twenty percent (roughly a million words) of our intended database parsed We hope to have the entire database parsed and indexed by late spring and to have some more precise assessments of the effectiveness of different indexing strategies and of the matching algorithms This will also provide an opportunity to experiment with the 'index and associate approach to phrasal venation discussed above introducing and then indexing to examples of semantically associated phrasal variations

We are also planning to apply the structures and algorithms described here to non-textual domains including the description of images In this domain individuals will correspond to salient blobs of color and texture with micro-relations describing their geometric and topological relationships to one another This may prove an additional challenge to our matching and indexing algorithms since the number of micro-relations will likely be quite large (compare to the text case where it is of the same order as the number of individual words)

References

[Falkenhainer et *al* 1989] Falkenhainer B Forbus K and Gentner D The Structure-Mapping Engine Algorithms and Examples *Artificial Intelligence* (41) 1989

[Gentner *et al* 1991] Gentner D and Forbus K MAC/FAC A model of similarity-based retrieval *Proceedings of the Thirteenth Annual Conference of the Cognitne Science Socity* Lawrence Earlbaum 1991

lHolyoak and Thagard 1989] Holyoak K and Thagard P A computational model of analogical problem solving In *Similarity and Analogical reasoning* edited by S Vosmadou and A Ortonv Cambridge University Press 1989

[Lakotf l987]Lakoff G *Women Fire and Dangerous Things* University of Chicago Press 1987

[Miller 1990] Miller G Wordnet An On-line Lexical Database *International Journal of Lexicgrophy* 3(4) 1990

[Mitchell, 199]] Mitchell M *Analogy making as perception,* MIT Press 1991

[Sleator and Temperly 1991]Sleator D and Temperly D , 'Parsing English with a Link Grammar Carnegie Mellon School of Computer Science technical report CMU-CS-91 196

[Thagard *et al* 1990] Thagard P Holyoak K Nelson G and Gochfeld D "Analog Retrieval by Constraint Satisfaction , *Artificial Intelligence* (46)

[Vorhees 1994] Vorhees, E M Query Expansion Using Lexical-Semantic Relations, in *Proceedings of SIGIR 94,* 1994 Croft B and Reisbergen C ed Spnnger-Verlag 1994