INTELLIGENCE · ARTIFICIAL AND COLLECTIVE INTELLIGENCE

Automated Entity Extraction and Relationship Mapping from Unstructured Data

Most of the information intelligence teams work with is unstructured text. This article explains how software identifies people, organisations and places in text, extracts the relationships between them, and builds maps that reveal connections no single document shows.

ModQu research article · 951 words · 11 references (APA 7th edition)

News articles, company filings, customer emails, interview notes and social media posts all contain valuable intelligence, but in a form computers cannot easily use. They are **unstructured**: free text rather than tidy rows and columns. **Information extraction** is the set of techniques that turn such text into structured data. Sarawagi (2008) describes it as the automatic extraction of structured information, such as entities, relationships and events, from unstructured sources. For intelligence work, two tasks are central: finding the entities mentioned in text, and identifying how they are connected.

Finding entities

Named entity recognition identifies mentions of specific things in text and classifies them by type: people, organisations, locations, dates, amounts of money and so on. Nadeau and Sekine (2007) review the development of the field from early systems based on hand-written rules and lists of names to statistical methods that learn from annotated examples. Government evaluation programmes helped drive progress; the Automatic Content Extraction programme, for instance, defined standard tasks and test data for recognising entities, relations and events (Doddington et al., 2004).

Modern systems use neural networks. Lample et al. (2016) showed that neural architectures could achieve high accuracy without hand-crafted features, by learning from the characters and words in context. Li et al. (2022) review the deep learning methods that now dominate the field. Widely used open-source toolkits, such as Stanford CoreNLP, package these capabilities so that analysts can apply them without building systems from scratch (Manning et al., 2014).

Linking mentions to real entities

Finding a name is only the first step. The same company may appear as "International Business Machines", "IBM" and "the company". Two different people may share a name. **Entity resolution** decides which mentions refer to the same real-world entity. Christen (2012) describes the methods: comparing names and attributes, measuring their similarity, and deciding when records match. Errors here are costly. Merging two different people creates false connections; failing to merge mentions of one person hides real ones.

Extracting relationships

Relation extraction identifies how entities are connected: who works for which organisation, which company owns another, who met whom and where. Early systems required many hand-labelled examples for each relation type. Mintz et al. (2009) introduced **distant supervision**, which uses an existing database of known facts to label training text automatically: if a database says a person founded a company, sentences mentioning both are treated as likely examples of the "founded" relation. This made it possible to learn many relation types at scale, at the cost of some noisy labels.

Building the map

Extracted entities and relations can be stored as a **knowledge graph**, in which entities are nodes and relations are labelled links between them. Hogan et al. (2022) describe how knowledge graphs represent information in this way and how they can be queried and reasoned over. Figure 1 shows a simple example built from three sentences.

From text to a relationship map

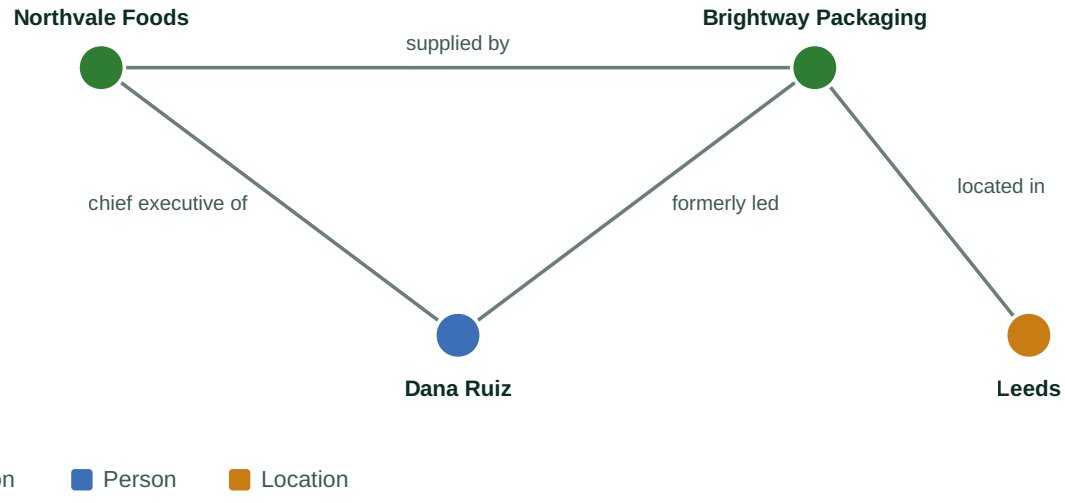
Entities and relations extracted from three fictional news snippets

"Northvale Foods appointed Dana Ruiz as chief executive."

"Dana Ruiz previously led Brightway Packaging in Leeds."

"Northvale Foods signed a supply deal with Brightway Packaging."

Extracted: ORG · PERSON · LOCATION and the relations between them



Fictional example by ModQu; method based on Sarawagi (2008) and Hogan et al. (2022)

ModQu.com

Figure 1. From text to a relationship map. Each snippet contributes one or two links; together they reveal that a new supplier is led by the chief executive's former company. Fictional example by ModQu; method based on Sarawagi (2008) and Hogan et al. (2022).

No single sentence in the example says that Northvale's new supplier is the chief executive's former company. The map makes that connection visible. This is the main value of relationship mapping: it combines fragments from many sources into a picture that shows indirect links, clusters and central figures.

Once a map exists, the methods of **social network analysis** can be applied. Wasserman and Faust (1994) set out the standard measures, such as how many connections each node has, which nodes lie on paths between others, and which groups are densely linked. Krebs (2002) showed how mapping publicly reported connections among the hijackers in the September 2001 attacks revealed the structure of the network after the event, and drew lessons about how such maps can and cannot be used.

Uses in business

Market and competitive intelligence teams use entity extraction and relationship mapping to:

- track executives moving between companies and the links they carry with them;
- map supply chains from news, trade data and filings;
- identify the partners, investors and customers of competitors;
- detect conflicts of interest or undisclosed related parties in due diligence;
- group customer complaints by product, location and issue.

For example, a procurement team screening a new supplier might extract every person and company named in two years of trade press coverage about it. The resulting map could show that the supplier shares a director with a firm the company had previously dropped for quality failures, a link that would take days to find by reading alone.

Limitations

Automated extraction makes errors. Names are missed or misclassified, especially unusual names and those in other languages. Relations are inferred wrongly from sentences that mention two entities without connecting them. Entity resolution merges or splits incorrectly. Errors compound as the map grows, and a false link can suggest a connection that does not exist. Krebs (2002) warned that network maps built from incomplete data can mislead, because missing nodes and links change the apparent structure.

Good practice includes checking a sample of extractions by hand, keeping a link from every node and relation back to the source text, recording confidence for each extracted fact, and treating surprising connections as leads to verify rather than findings.

Conclusion

Entity extraction and relationship mapping turn large volumes of text into structured knowledge about who and what is connected. They let intelligence teams see indirect links and patterns that no single document reveals. Their output is only as reliable as the extraction and matching behind it, so every map should remain traceable to its sources and every important connection should be confirmed before it informs a decision.

References

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer. https://doi.org/10.1007/978-3-642-31164-2

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004). The Automatic Content Extraction (ACE) program—Tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (pp. 837–840).

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2022). Knowledge graphs. *ACM Computing Surveys*, *54*(4), Article 71. https://doi.org/10.1145/3447772

Krebs, V. E. (2002). Mapping networks of terrorist cells. *Connections*, *24*(3), 43–52.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 260–270). https://doi.org/10.18653/v1/N16-1030

Li, J., Sun, A., Han, J., & Li, C. (2022). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50–70. https://doi.org/10.1109/TKDE.2020.2981314

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). https://doi.org/10.3115/v1/P14-5010

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing* (pp. 1003–1011).

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26. https://doi.org/10.1075/li.30.1.03nad

Sarawagi, S. (2008). Information extraction. *Foundations and Trends in Databases*, *1*(3), 261–377. https://doi.org/10.1561/1900000003

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.