

埋め込み表現を活用した
OKR の Objective と Key Results の定量的整合性分析
A Quantitative Consistency Analysis of Objectives and Key Results
in OKRs Using Embedding Representations

江刺 裕太[†], 松本 萌[†], 三浦 紘嵩^{†,‡}, 大澤 正彦[†]
Yuta Esashi, Moe Matsuki, Hirotaka Miura, Masahiko Osawa

[†] 日本大学文理学部次世代社会研究センター, [‡] 高千穂大学商学部

[†] RINGS, College of Humanities and Sciences, Nihon University, [‡] Faculty of Commerce, Takachiho University
chyu25006@g.nihon-u.ac.jp

概要

本研究は、Objectives and Key Results における Objective (O) と Key Results (KR) の整合性を定量評価する手法を検討する。KR 全ての達成が O の達成になるという前提を、ベクトル化した埋め込み表現を用いて「KR の総和は O と類似する」ことを検証する。23 件の OKR データを用いて対応しない O と KR の総和より対応する KR の総和の方が類似することが検定により確認できた。

キーワード: Objectives and Key Results, 埋め込み表現, 類似度, 意味的一貫性, 定量評価

1. はじめに

OKR (Objectives and Key Results) は、目標 (Objective) と主要な成果指標 (Key Results) をセットで管理するフレームワークであり、近年、個人・チーム・組織の目標設定に広く採用されている。OKR はシリコンバレーの企業で発展し、Google の導入を機に広く知られるようになった。現在では成長戦略やチーム強化の手法として多くの企業が採用している。

OKR は、定性的な目標である Objective に対し、2～5 個の定量的な成果目標である Key Results を整合させるのが基本であり、「すべての Key Results を達成することで Objective も達成される」ことが前提とされている (Doerr, 2018)。しかし、定性的な Objective と定量的な Key Results を整合させることが実務上困難であるという課題が報告されている (Stray et al., 2022)。こうした背景を受けて、Objective と Key Results の整合性を担保するために、他の目標管理手法との統合を試みる研究も存在する (Trinkenreich et al., 2019)。これらの背景から、Objective と Key Results の整合性をより客観的かつ定量的に評価する手法の必要性が高まっている。

本研究では、文章の意味をベクトルで表現する埋め込み表現を用いることで定量的に評価する方法を検討

する。定量的に評価できることで、整合性の判断を再現性のある形で行えるようになる。

2. 関連研究

2.1 Objectives and Key Results (OKR)

OKR は、組織や個人が達成すべき目標と、その達成度を評価するための主要な成果指標をセットで定義する目標管理のフレームワークである (Doerr, 2018)。

OKR は以下の 2 つの要素から構成される：

- Objective：達成したい目標を定性的に示す。方向性や意図を端的に表現するもの。
- Key Results：Objective の達成度を評価するための定量的な主要な成果指標。

たとえば、以下はコンサルティングサイト Workpath(2024) で公開されている「E-Commerce OKRs」に基づく、カスタマーサポートの質向上を目指す OKR の一例である：

- **Objective**：Increase the on-time delivery rate for E-commerce Logistics, ensuring customer satisfaction and trust in our services
- **Key Results1**：Improve Package Tracking Accuracy: 95% of all tracking updates are correct
- **Key Results2**：Reduce Delivery Errors: Error rate below 2% for all deliveries
- **Key Results3**：Boost Delivery Speed: Average delivery time under 24 hours
- **Key Results4**：Enhance Customer Support: 90% positive feedback on support interactions
- **Key Results5**：Optimize Routing Efficiency: 85% of routes use optimal paths

この例では、「E コマース物流における納期遵守率の向上」という定性的な目標 (Objective) に対して、追跡精度や配送エラー率、平均納品時間、顧客対応の評価、ルート最適化率などの測定可能な項目が Key Results として紐づけられている。Objective が「何をを目指す

か」を示す一方で、Key Results はその達成状況を測る具体的な手段を提供している。すなわち、Objective は複数の Key Results によって意味的に構成される全体として位置付けられる。つまり、Objective とそれぞれの Key Results は Part-of 関係であると言える。

2.2 埋め込み表現

自然言語処理における埋め込み表現は、単語や文章の意味をベクトルで表現する手法であり、語彙や文の意味的関係を計算的に扱うことを可能にする。例えば、king - man + woman $\approx$ queen のような意味的ベクトル演算が可能である (Mikolov et al., 2013)。

教育分野では、米国全体の教育基準と各州の教育基準の一貫性を調査するために、教育基準の文章を埋め込み表現として数値ベクトル化し、基準間の類似度を測定する試みもなされている (Butterfuss & Doran, 2024)。

3. 仮説

OKR は「すべての Key Results を達成することで Objective も達成される」という前提に基づき、Objective は Key Results の集合として捉えられる。本研究では次の仮説を立てる (図 1)。

仮説 適切に設定された OKR では、Objective の埋め込み表現は対応する Key Results の埋め込み表現の総和で近似可能である。

任意の Objective x と任意の Key Results の集合 y の埋め込み表現間のコサイン類似度を計算する関数を

$$f(x, y) = \cos\left(\vec{W}_o^x, \sum_{i=1}^n \vec{W}_{kr_i}^y\right)$$

と定義する。OKR の集合を D とし、任意の OKR を $target \in D$ とする。このとき、仮説は以下のように表される：

$$\forall target \in D, \quad f(target, target) \approx 1 \quad (1)$$

4. 検証方法

本稿では、3. 章で提示した仮説を検証するため、企業向けコンサルティングサイト Workpath (2024) で公開されている OKR データを用い、以下の関係式が成り立つかを確認する (図 2)。

$$\forall target \in D, \forall other \in D \setminus \{target\}, \\ f(target, target) > f(target, other) \quad (2)$$

$other$ は $target$ とは異なる D 内の別の OKR 要素を表す。式 2 は、任意の OKR $target$ において、その Objective と対応する Key Results の総和とのコサイン類似度 ($f(target, target)$) が、他の OKR に属す

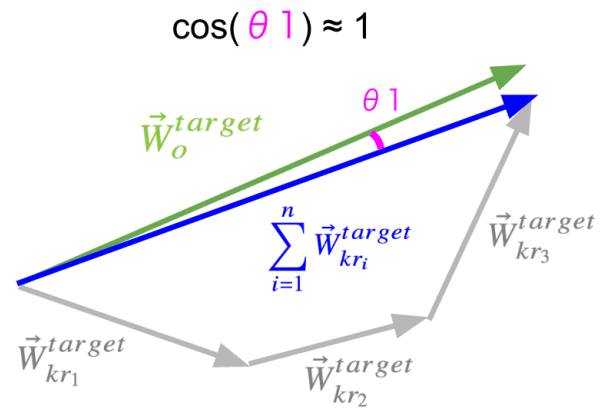


図 1 Objective と Key Results の関係仮説: 適切に設定されている OKR において $\vec{W}_o^{target}$ (target の Objective の埋め込み表現) は $\sum \vec{W}_{kr}^{target}$ (target の Key Results の埋め込み表現の総和) とのコサイン類似度が 1 に近似する

る対応しない Key Results の総和とのコサイン類似度 ($f(target, other)$) よりも高くなることを示す。この関係が統計的に広く成り立つことが確認できれば、OKR の整合性を埋め込み表現を用いて定量的に評価できる可能性があるといえる。

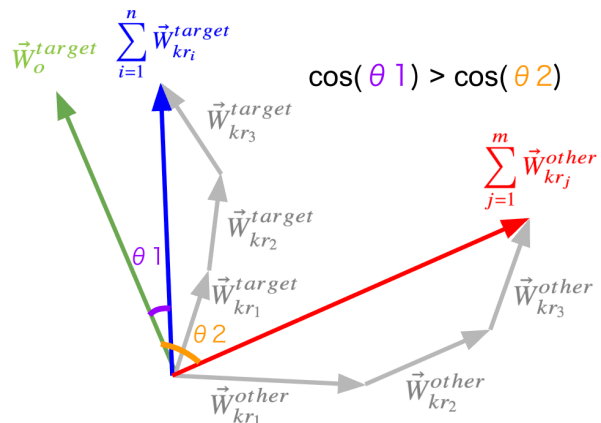


図 2 検証のための仮説: $f(target, target)$ は $f(target, other)$ より大きい

4.1 OKR データ

本実験では、コンサルティングサイト Workpath(2024) で公開されている 23 件の OKR サンプルデータを用いた。Workpath は、専門的知見を有する企業向けコンサルティング会社であり、本研究では整合性の高い OKR データとして同サイトのデータを採用した。OKR は業種や生活の分野を想定して作成されたもので、多様な業界 (金融, IT, ヘルスケア, 非営利組織, スタートアップなど) や個人向け目標を幅広くカバーしている。

4.2 前処理方法

仮説の検証を行うために、OKR の自然言語を埋め込み表現に変換する。変換手段として、OpenAI が提供する text-embedding-ada-002 を使用する。このモデルは、任意の長さのテキストを 1536 次元の高次元ベクトルに変換する。この高次元のベクトルは意味的な情報を保持した埋め込み表現である。Transformer アーキテクチャに基づき、検索、分類、意味的類似度評価など多様な NLP タスクで活用されている (OpenAI,2023)。

4.3 定義

各 OKR に対して 1 つの $f(\text{target}, \text{target})$ を算出できるため、 $f(\text{target}, \text{target})$ は合計 23 件となる。一方で、各 OKR (全 23 件) に対し、それ以外の 22 件の OKR を対応しない Key Results 群として組み合わせることができるため、 $f(\text{target}, \text{other})$ は $23 \times 22 = 506$ 件得られた。以後、それぞれのサンプル集合を関係群と無関係群と定義する。

本検証では、OKR の整合性があるとされるペア関係群と、整合性がないとされるペア無関係群の間で、コサイン類似度に有意な差が存在するかを評価する。具体的には、無関係群よりも高い類似度を示す傾向があるかどうかを検討する。

5. 検証結果

5.1 統計的分析結果

関係群と無関係群を比較するために箱ひげ図を並べてプロットした (図 3)。その結果、関係群の方が、無関係群より高い位置にプロットされていることが確認できた。また、関係群と無関係群の平均値 (±標準偏差) はそれぞれ 0.8892 (± 0.0141) と 0.8266 (± 0.0223) であった。両者の間に有意な差が存在するかを確認するため、各データが正規分布に従うかを検定した。シャピロ・ウィルク検定 ($p < 0.05$) の結果、関係群の p 値は 0.0481 であり、正規分布に従わなかった。一方、無関係群の p 値は 0.2740 であり、正規分布に従った。このように群によって正規性の有無が異なるため、ノンパラメトリックな検定手法を用いた。本実験では、関係群と無関係群のサンプルサイズは異なり、データ間に対応関係は存在しない、したがって、マンホイットニーの U 検定 ($p < 0.05$) を採用した。検定の結果、 $p = 1.38 \times 10^{-15}$ であり、関係群と無関係群の間に有意な差が確認された。

5.2 ヒストグラムによる可視化の結果

次にヒストグラムによる可視化を行なった (図 4)。比較のため、関係群を赤色、無関係群を青色で図示した。図から、全体としては正規分布に近い形状を示し、

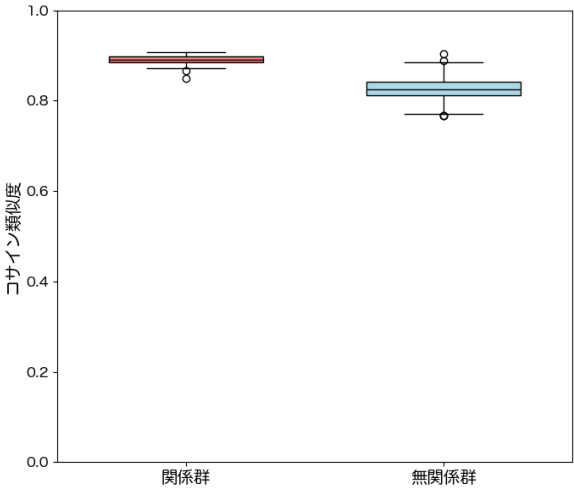


図 3 関係群と無関係群を比較した箱ひげ図。左が関係群、右が無関係群の箱ひげ図。縦軸はコサイン類似度の値を示す。

その中で赤色のサンプルが右側に分布し、他の領域は青色のサンプルが中心であることが確認できた。このことから、関係群はサンプル全体に対して相対的に高い値を示すことがわかった。

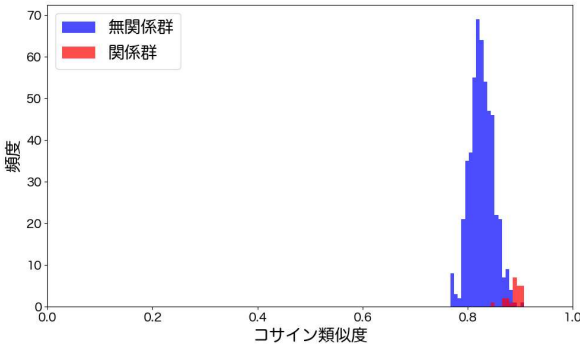


図 4 関係群 (赤色) と無関係群 (青色) のヒストグラムを用いた比較可視化。

6. 考察

6.1 統計的傾向の検証の考察

関係群 と 無関係群 の間に有意な差が認められ、さらにヒストグラムからも 関係群 が全体の分布に対して高い値を示すことが確認された。この結果から、適切に設定された OKR において式 2 が支持される可能性が示唆されたと考えられる。

6.2 閾値の算出

式 2 の関係を基に、「適切に設定された OKR」を定量的に識別するための閾値の導出を試みた。具体的には、関係群 をポジティブクラス、無関係群 をネガティブクラスとし、ROC 曲線を描いて Youden' s J 指数により最適閾値を導出した (図 5)。その結果、コサイン類似度 0.8656 を閾値とした場合、感度 95.65%、偽陽

性率 4.35% の分類性能が得られた。

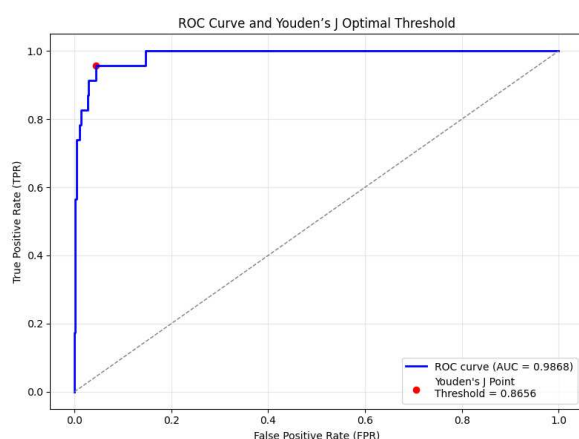


図 5 関係群と無関係群の ROC 曲線と Youden's J 指数. 閾値 0.8656 の場合、感度 95.65%, 偽陽性率 4.35% の分類性能となる (赤い点)。

6.3 閾値に対する考察

閾値に対して反証された 4.35% の事例について詳細に分析する。反証されたものは主に 2 種類存在し、無関係群の中で閾値 (0.8656) を上回ったものが 22 件、 $\mathcal{D}_{\text{corr}}$ の中で閾値を下回ったものが 1 件確認された。前者の反証の例の中で、「E-Commerce OKRs」の Objective と「Operations OKRs」の Key Results の総和とのコサイン類似度が 0.90 と非常に高く、反例の中で最も高い値を示していた。この現象をより詳細に把握するため、「E-Commerce OKRs」の Objective に対し、対応する Key Results および対応しない Key Results とのコサイン類似度を総和ではなく各 Key Results に対し算出したところ、対応しない Key Results との類似度が平均的に高い類似度を示す結果となった (対応しない Key Results 群の平均 0.8555, 標準偏差 0.0001, 対応 Key Results 群の平均 0.8261, 標準偏差 0.0182)。このことから反証した原因は、対応しない Key Results であっても Objective と高い意味的整合性を持っていた場合は、Key Results ベクトルの総和においても高い類似度となったからと考えられる。

しかし反証した結果が必ずしも間違いと言い切ることは難しい。特に今回の反証したサンプルの妥当性は、専門的な知識が必要であるため、これ以上の評価は困難であった。つまり、閾値 0.8656 を基準として設定したことに対する実用的妥当性や解釈可能性といった定性的な観点での評価は困難であった。このため、今後は被験者実証を通じて、ユーザの評価に基づいて本稿の閾値による評価の妥当性を検証することが重要であると考えられる。

後者の事例では、「Key Company OKRs」の Objective と対応する Key Results の総和とのコサイン類似度が

0.85 であり、閾値をわずかに下回っていた。ただし、関係群において閾値未満となったのはこの 1 件のみであり、差分も 0.01 程度にとどまる。この結果は、統計的ばらつきや計測誤差の範囲内である可能性も考慮されるべきであり、仮説全体の傾向に対する反証とは言い難い。

7. まとめ

本研究では、OKR (Objectives and Key Results) における Objective と Key Results の整合性を評価するために、「Objective の埋め込み表現は対応する Key Results の埋め込み表現の総和と高いコサイン類似度を持つ」と仮説を立て、Objective が対応しない Key Results 群より対応する Key Results 群と高い類似度を示すことを検証した。その結果、仮説について統計的に有意な差が確認された。さらに、ROC 分析を通じて整合性評価に活用可能な閾値を検討し、反例となる事例の分析をした。その結果、Objective と各 Key Results の埋め込み表現が類似する場合、反証する結果になる可能性が示唆された。しかし、今回の反証した結果が実際の OKR における整合性評価の判断基準となり得るかどうかは、専門的知識や個人の目標設定における意図に依存するため、本研究では最終的な判断には至らなかった。今後は、提示した閾値の有効性を確かめるため、被験者実験を通じた実用性の検証が求められる。

文 献

- Doerr, J. (2018). *Measure what matters*. Penguin Random House.
- Stray, V., Gundelsby, J. H., Ulfsnes, R., & Moe, N. B. (2022). How agile teams make Objectives and Key Results (OKRs) work. *Proceedings of the International Conference on Software and System Processes and International Conference on Global Software Engineering (ICSSP '22)*, 104–109. Association for Computing Machinery. <https://doi.org/10.1145/3529320.3529332>
- Trinkenreich, B., Santos, G., Barcellos, M., & Conte, T. (2019). Combining GQM+Strategies and OKR: Preliminary results from a participative case study in industry.
- Workpath. (2024). *OKR Examples*. Retrieved from <https://www.workpath.com/en/okr-examples>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. <https://arxiv.org/abs/1301.3781>
- OpenAI. (2023). *OpenAI platform documentation: Embeddings*. Retrieved from <https://platform.openai.com/docs/guides/embeddings>
- Butterfuss, R., & Doran, H. (2024). An application of text embeddings to support alignment of educational content standards. *Educational Measurement: Issues and Practice*, 44(1), 73–83. <https://doi.org/10.1111/emip.12641>